STUDIES OF MOUSE ACTIN

GENOMIC CLONES

Carolyn E. Begg

Thesis submitted to the University of Glasgow

for the degree of Doctor of Philosophy

Department of Biochemistry

February, 1987

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

To

My Parents

Acknowledgements

I would like to express my grateful thanks to Dr D.P. Leader for his excellent help and supervision throughout the course of this project.

I would like to thank Professor R.M.S. Smellie for making the facilities of the Department of Biochemistry available for this research.

My thanks are also extended to my colleagues in C36, past and present, for their friendship and for making the lab such an enjoyable place to work. My special thanks go to Irene Gall, for showing extreme patience while passing on her technical expertise and for showing me great kindness since arriving in Glasgow. Special thanks also goes to my great friend and colleague Ying Min Man, for always being extremely kind and helpful, and for introducing me to Chinese cuisine. I am also extremely grateful to all my friends in the Department, especially my neighbours in lab C35, who helped me through my good and bad days.

I would like to express my grateful thanks to Dr Bob Eason for his advice and encouragement in helping me to understand a little of the mysteries of computers.

The greatest thanks of all is to my parents, for the endless support and encouragement they have given me throughout my years at university.

Abbreviations

The abbreviations recommended by the *Biochemical Journal* in its Instructions to Authors (*Biochemical Journal* (1985) **225**, 1 - 26) have been used throughout this thesis with the following additions :

bp	base pairs
BSA	bovine serum albumin
cpm	counts per minute
DNase	deoxyribonuclease
dNTP	deoxynucleotide-5'-triphosphate
EMBL	European Molecular Biology Laboratory
kb	kilobase, (1000 base pairs)
LINE	long interspersed repetitive elements (L1 elements)
L1Md	L1 is followed by a two-letter genus and species designation, such as L1Md for the L1 family in *Mus domesticus*
MY	million years
PEG	polyethylene glycol
pfu	plaque forming units
RNase	ribonuclease
SDS	sodium dodecyl sulphate
SINE	short repetitive interspersed elements

Contents

Page

Page

Page

Lists of figures

Page

Chapter 3 :

Page

Page

Page

Page

List of Tables

Summary

This thesis describes studies of two genomic clones λmA14 and λmA36, which had been isolated from a mouse genomic lambda library using a rat skeletal muscle actin cDNA probe, and which electron microscopic heteroduplex analysis had shown to contain a similar, although not identical self-hybridising (foldback) structure adjacent to the actin-like region. The objective of these studies was to determine the extent of the similarity between λmA14 and λmA36, and the nature of the actin-like DNAs and the DNA constituting the foldback structures.

Detailed restriction maps of λmA14 and λmA36 were constructed in order to compare these clones. This was achieved by a combination of the following techniques : (i) single restriction enzyme digestion, (ii) hybridisation of a ^{32}P-labelled actin probe to the products of single and double restriction enzyme digestion, (iii) partial restriction enzyme digestion followed by hybridisation to a ^{32}P-labelled oligonucleotide complementary to the cohesive end of the short arm of bacteriophage lambda, (iv) generation of subclones covering most of the mouse DNA inserts in λmA14 and λmA36, and subjecting these to single and double restriction enzyme digestion. The resulting maps showed that over a region of 11.0kb there were 25 restriction endonuclease sites which appeared to be identical in the two clones and 11 which were clearly different, after allowing for an extra inserted 0.5kb of DNA in λmA36 that was also found by electron microscopy. This suggested that clones λmA14 and λmA36 contain at least 11.0kb of similar but not identical DNA, and this suggestion was supported by the positive cross hybridisation of

fragments from the two clones and partial nucleotide sequence determination of the DNA near the left and right-hand extremities of the apparent similarity. Comparison of these and other sequences from λmA14 and λmA36 indicated an average difference of 5.7%. This suggests that the two sequences diverged from a common ancestor 2.6 MY ago.

Partial nucleotide sequencing was used to determine the nature of the actin-like DNA in clones λmA14 and λmA36. The portion of actin-like DNA sequenced in λmA14 corresponds to that specifying amino acids 1 to 302. Predicted amino acids at the N-terminal end of this sequence identified this as being related to the γ-cytoplasmic member of the six mammalian isoforms of actin. The partial sequence of the actin-like gene of λmA36 showed it to be related to a cytoplasmic β- or γ-actin, although lack of sequence at the N-terminal end prevented more precise identification.

The actin-like gene of λmA14 contained a significant number of differences in predicted amino acid sequence from γ-actin, and several termination codons. Furthermore it lacked introns. These features indicate that λmA14 contains an actin pseudogene of the processed type. This also appeared to be truncated at its 5' end. Comparison of the nucleotide sequence with that of a mouse γ-actin cDNA clone showed 5% difference, suggesting a relatively recent origin.

As λmA14 and λmA36 had similar restriction maps over much of the foldback region, the structure of this foldback was analysed in the single clone, λmA14. Areas of the three subclones thought to contain the stem of the foldback structure were sequenced, and homologous regions were identified in each subclone, that could account for the electron microscopic features.

These were a region of at least 1.5 kb, adjacent to the actin, orientated in one direction (designated LH) and two regions of 1.3 kb and at least 1.0 kb, respectively (RH1 and RH2) orientated in the opposite direction. The sequence of the two regions RH1 and RH2 had an overlap of approximately 460bp. The region RH1 is outwith the DNA included in the smaller clone, λmA36, and this and the overlap of RH1 and RH2 adequately account for the electron microscopic differences of λmA14 and λmA36 in regions where they have similar restriction maps.

To determine the nature of the sequences constituting the stem of the foldback element a ^{32}P-labelled fragment of this DNA was hybridised to digested mouse chromosomal DNA subjected to agarose gel electrophoresis and transferred to nitrocellulose. The strength of the hybridisation indicated that the stem sequence was repetitive and, against a background smear, discrete bands were observed, the length of which were similar to those of the previously characterised L1Md, mouse middle repetitive DNA family. The sequences of the foldback area of λmA14 were compared to that of a recently published 'full-length' L1Md DNA sequence, confirming that the stem DNA of the foldback loop is composed of L1Md sequence. The foldback structure in λmA14 is composed of specific regions of three L1Md LINE members. One L1Md member (L1Md-LH), was contiguous with the truncated 3' end of the actin pseudogene of λmA14 and formed the left-hand arm of the stem. The right-hand arm was formed from two L1Md members (L1Md-RH1 and L1Md-RH2), which are located approximately 5.2 and 11.0kb respectively to the right, of the left-hand member, in the opposite orientation.

The left-hand L1Md member is at least 3.3kb in length with its 5' end contiguous with actin DNA at a position approximately 100bp from its expected 3' end. The sequence of the 3' end of the left-hand L1Md member

was not determined but hybridisation with a probe containing the extreme 3' end of a different γ-actin sequence, located the displaced 3' end of the actin pseudogene to a particular subclone, at least 3.1kb from the 5' end of L1Md-LH. Thus L1Md-LH has inserted independently into the γ-actin pseudogene. This measurement, together with the known length of a complete L1Md member and the presence of an internal deletion of 2.4kb in L1Md-LH, indicated that L1Md-LH most likely contains intact 5' and 3' ends. L1Md-RH2 appears to be truncated at both ends, whereas L1Md-RH1 is truncated at only the 3' end. L1Md-RH1 and L1Md-LH possess several features in common which differ from the prototype full-length L1Md member. These include the same 5' end containing 1 $^{2}/_{3}$ copies of a 208bp tandem repeat, and a common 42bp insertion, and suggest the possibility of gene correction at some stage of their existence.

The results presented do not allow an unequivocal decision as to whether the similar regions in λmA14 and λmA36 are the result of a gene duplication or amplification event, although indirect considerations favour the former possibility. However it is possible that the L1Md members identified in this work played a role in the original duplication or amplification of this large region of the mouse genome.

CHAPTER 1 Introduction

1.1 Actin proteins and genes

1.1.1 Actin proteins

Actins are highly conserved proteins which are found ubiquitiously in eukaryotic cells. Amino acid sequence data has demonstrated the presence of several distinct actin isotypes in vertebrates, and these isotypes can generally be classified as either 'cytoplasmic' or 'muscle' actins (Vandekerckhove & Weber 1978a). Cytoplasmic actins are found in non-muscle cells, where they are utilised to form the cellular microfilaments which function in cell motility and mitosis (Vandekerckhove & Weber, 1978a). The number of cytoplasmic isoforms ranges from at least two in mammals (Vandekerckhove & Weber, 1978a) to three, or even more, in birds and amphibians (Vandekerckhove et al., 1981; Bergsma *et al.*, 1985). Muscle actins are essential components of the contractile apparatus of muscle cells and are subdivided into either striated or smooth isoforms, according to the muscle cell type in which they predominate. The striated muscle isoforms may be coexpressed in a tissue under at least some circumstances (Gunning *et al.*, 1983b; Hayward & Schwartz, 1986) with α-skeletal muscle actin representing the predominant form in adult skeletal muscle and α-cardiac muscle actin prevailing in adult cardiac tissue (Vandekerckhove & Weber, 1978b; 1979a). The smooth muscle actins appear to be similarly coexpressed (Vandekerckhove et al., 1981). In the genital and gastrointestinal tracts, γ-smooth muscle actin

predominates, while in vascular tissue, such as aorta, α-smooth muscle actin is the primary isotype (Vandekerckhove & Weber, 1979a; 1984; Gabbiani *et al.*, 1981).

Only limited differences in amino acid sequence exist between the actin isotypes of vertebrates, and these are located primarily in the amino terminal region. Table 1.1 shows the positions in the amino acid sequence at which differences occur between the six actin isoforms of mammals. There are 4-6 amino acid replacements between the different muscle types; 4 amino acid replacements between the two cytoplasmic actins and 25 amino acid replacements between the cytoplasmic and skeletal muscle actins (Vandekerckhove & Weber, 1979a). Actins from diverse organisms are extremely similar. For example, chicken, bovine and rabbit skeletal muscle actins have identical amino acid sequences (Vandekerckhove and Weber, 1979a,b), which differ from the yeast actin sequence at only 49 out of 375 positions (Gallwitz & Sures, 1980; Ng & Abelson, 1980).

All eukaryotes synthesize one or more cytoplasmic actins isoforms (Vandekerckhove et al., 1981). The vertebrate non-muscle β and γ-actins are considered functionally and evolutionarily more closely related to the actins found in the lower, unicellular, eukaryotes. In *Drosophila melanogaster,* actins with amino acid sequences resembling those of the vertebrate cytoplasmic actins are utilised to form the actin filaments of sarcomeric muscle (Fyrberg *et al.*, 1981). It has been proposed that during early chordate evolution a novel actin isoform arose and now functions in the sarcomeres of muscle cells (Vandekerckhove *et al.*, 1983). In the time prior to the divergence of mammals and birds, this gene apparently underwent two successive rounds of duplication to produce the four muscle-actin isoforms found in mammals and birds today (Vandekerckhove *et al.*, 1983). Thus the muscle-actin isotypes must have been under strong

Table 1.1 Differences in the amino acid sequences of the actin isoforms

Residue number	Actin types					
	Skeletal muscle	Cardiac muscle	Smooth muscle (stomach)	Smooth muscle (aorta)	Non-muscle β-type	Non-muscle γ-type
1	Asp	Asp	-	Glu	Met	-
2	Glu	Asp	Glu	Glu	Asp	Glu
3	Asp	Glu	Glu	Glu	Asp	Glu
4	Glu	Glu	Glu	Asp	Asp	Glu
5	Thr	Thr	Thr	Ser	Ile	
6	Thr	Thr	Thr	Thr	Ala	
10	Cys	Cys	Cys	Cys	Val	Ile
16	Leu	Leu	Leu	Leu	Met	
17	Val	Val	Cys	Cys	Cys	
76	Ile	Ile	Ile	Ile	Val	
89	Thr	Thr	Ser	Ser	Thr	
103	Thr	Thr	Thr	Thr	Val	
129	Val	Val	Val	Val	Thr	
153	Leu	Leu	Leu	Leu	Met	
162	Asn	Asn	Asn	Asn	Thr	
176	Met	Met	Met	Met	Leu	
201	Val	Val	Val	Val	Thr	
225	Asn	Asn	Asn	Asn	Gln	
259	Thr	Thr	Thr	Thr	Ala	
266	Ile	Ile	Ile	Ile	Leu	
271	Ala	Ala	Ala	Ala	Cys	
278	Tyr	Tyr	Tyr	Tyr	Phe	
286	Ile	Ile	Ile	Ile	Val	
296	Asn	Asn	Asn	Asn	Thr	
298	Met	Leu	Leu	Leu	Leu	
357	Thr	Ser	Ser	Ser	Ser	
364	Ala	Ala	Ala	Ala	Ser	

The table indicates the positions in the amino acid sequence at which exchanges have been detected between the different actin isoforms. Positioning of the amino acids in the actin sequence is made in analogy to rabbit skeletal muscle actin (Collins & Elzinga, 1975; Lu & Elzinga, 1977; Vandekerckhove & Weber, 1978c). Amino acid residues in which the four muscle actins differ among themselves are underlined.

selective pressure to maintain their amino acid sequence since they arose.

1.1.2 Actin genes

The isolation of actin cDNA clones (Ponte *et al.*, 1983; Gunning *et al.*, 1983b) allowed the number of actin-related sequences in the genome of different organisms to be determined by hybridisation and to be isolated. individual genomic sequences. The structural characterisation of these sequences (which will be referred to loosely as actin 'genes'), has revealed a number of interesting features which are discussed in the sections below:

(a) Gene number

When the genomic DNA of an organism is analysed by Southern blotting to an actin probe under low stringency washing conditions, the recognisable actin genes of the organism are revealed. It appeared from such genomic blots that the number of actin genes in higher eukaryotes varies widely. For example, chicken contains 4 - 7 actin genes (Cleveland *et al.*, 1980), human DNA 20 - 30 actin genes (Moos & Gallwitz 1982; Engel *et al.*, 1981), mouse DNA greater than 20 actin genes (Minty *et al.*, 1983) and rat 12 or more actin genes (Nudel *et al.*, 1982a). These numerous actin sequences are dispersed on different chromosomes throughout the mammalian genome (Soriano *et al.*, 1983). The number of actin genes in lower eukaryotes is also found to differ from one organism to another, *Drosophila melanogaster* contains 6 actin genes (Fyrberg *et al.*, 1981), yeast 1 actin gene (Gallwitz & Sures, 1980; Ng & Abelson, 1980), Dictyostelium 17 actin genes (McKeown & Firtel, 1981) and sea urchin 11 actin genes (Scheller *et*

al., 1981). In the lower eukaryotes the number of genes is roughly equivalent to the number of identified actin isoforms, however in the mammalian genome there is a much higher number of actin-related sequences than known actin isoforms.

Under high stringency washing conditions, only the most homologous sequence(s) remain hybridised to the genomic DNA and usually these correspond to the functional gene(s), (Minty *et al.*, 1983; Robert *et al.*, 1984; Weydert *et al.*, 1983). In this way it was possible to examine the number of genes coding for each isotype. Each actin isoform, like most structural proteins appears to be present in one copy per haploid genome (Minty *et al.*, 1983; Ponte *et al.*, 1983; Robert *et al.*, 1984). Many of the numerous actin sequences detected at low stringency in the mammalian genome were identified as dispersed processed pseudogenes (see section 1.2), derived from β or γ-actin mRNAs (Minty *et al.*, 1983; Carmon *et al.*, 1982). The extent to which these sequences have diverged from the actin coding sequence and hence, the time which has elapsed since their integration varies. The observation that the cytoplasmic actin genes but not the sarcomeric actin genes, are associated with the pseudogene families, has suggested a link in the expression of a gene in the germline cell to the production of large processed pseudogene families (Ponte *et al.*, 1983; see section 1.2). The high number of actin - related sequences is apparently restricted to the mammalian genome; in birds (Cleveland *et al.*, 1980) or in *Drosophila melanogaster* (Fyrberg *et al.*, 1980) for example, the number of genomic sequences corresponds to the number of known actin proteins.

(b) Gene structure

The high degree of sequence conservation between the actin proteins from a wide variety of organisms argues strongly that this multigene family arose by duplication and subsequent divergence from a common ancestral gene. In the course of evolution, certain regulatory and structural features of the loci have diversified to produce the specialised genes present today. Several representatives of the vertebrate striated (Fornwald *et al.*, 1982; Hamada *et al.*, 1982; Zakut *et al.*, 1982; Chang *et al.*, 1985; Eldridge *et al.*, 1985; Hu *et al.*, 1986), cytoplasmic (Bergsma *et al.*, 1985; Kost *et al.*, 1983; Nudel *et al.*, 1983; Ng *et al.*, 1985), and smooth muscle (Ueyama *et al.*, 1984; Carroll *et al.*, 1986; Chang *et al.*, 1984), actin gene subfamilies have been structurally characterized.

Each actin isoform is most likely encoded by a single gene (Minty *et al.*, 1983; Ponte *et al.*, 1983), which is not genetically linked to loci encoding either other members of the actin family (Czosnek *et al.*, 1983; Minty *et al.*, 1983; Gunning *et al.*, 1984a) or other contractile proteins (Czosnek *et al.*, 1982; Roberts *et al.*, 1985).

Due to the great conservation of the amino acid sequence among the actins, the nucleotide sequences of the coding regions of actin genes are highly conserved. When non-homologous actin isotypes are compared between species, the 5' and 3' non-coding regions of actin genes can be quite diverged, showing great variabilty in length and nucleotide sequence. On the other hand, comparison of homologous actin isoforms between species, shows a considerable degree of homology even between the untranslated portions of the mRNA. In birds and mammals, it has been demonstrated that the 3' untranslated region of actin mRNAs are unique to each actin isotype (Cleveland *et al.*, 1980; Minty *et al.*, 1981; Ponte *et al.*,

1983). The 3' untranslated regions of the human skeletal, cardiac, β and γ-actin mRNAs are capable of hybridising to the corresponding gene sequences of rodents (Ponte *et al.*, 1983, 1984). The 3' untranslated regions of rat (Mayer *et al.*, 1984) and human (Hamada *et al.*, 1982) cardiac actin genes show a high degree of homology; two-thirds of the 3' part of these regions exhibits 92.5% homology and the 5' part of this region exhibits 85% homology. However it appears that only the 3' untranslated region of the α-smooth muscle actin gene does not demonstrate this extensive evolutionary conservation (Carroll *et al.*, 1986), observed in the 3' untranslated region of α-skeletal (Hu *et al.*, 1986; Yaffe *et al.*, 1985; Gunning *et al.*, 1984b; Ordahl & Cooper, 1983), α-cardiac (Chang *et al.*, 1985; Eldridge *et al.*, 1985) and β-cytoplasmic actin genes (Yaffe *et al.*, 1985; Ponte *et al.*, 1984). The biological significance of the 3' untranslated conservation in these genes is unclear and therefore it is difficult to make an assessment of the significance of a lack of such conservation in the 3' untranslated region of the α-smooth muscle actin gene. Comparison of the 5' untranslated region of the human (Ponte *et al.*, 1984) and rat β-actin gene (Nudel *et al.*, 1983), revealed 80% homology, suggesting considerable conservation of this region of the gene.

Recently it was reported that three additional non-coding regions of the human β-actin gene are also highly conserved, including segments of the 5' flanking region, and two intervening sequences (Ng *et al.*, 1985). In all of the muscle actin genes examined thus far, TATA and CAAT boxes were located immediately upstream from the mRNA cap site, at the expected locations of -30 and -70, respectively (Carroll *et al.*, 1986; Nakajima-Iijima *et al.*, 1985). However in unicellular organisms, although these boxes occur,

they are not always at the expected location (Buckingham & Minty, 1983; Buckingham, 1985).

Structural characterisation of representative genes from several vertebrate multigene families has led to the observation that, in many cases, intron positions but not necessarily sequences are conserved (Breathnach *et al.*, 1981). However examination of actin genes revealed that although intron positions are somewhat conserved in deuterostomes (Fornwald *et al.*, 1982; Zakut *et al.*, 1982; see Table 1.2), such conservation is much less apparent in protosomes (Fyrberg *et al.*, 1981). These observations have led to much disagreement about whether the intron positions found in modern actin genes are the result of (a) the loss of some introns from a common ancestral actin gene which originally had many introns, (b) insertion of new introns into an intronless primordial actin gene, or (c) some combination of intron insertion or deletion. A comparison of the intron positions in the actin genes of deuterostomes to those found in the recently sequenced α-smooth muscle actin gene (Carroll *et al.*, 1986), sheds new light on this controversy. It was demonstrated that the structural sequence of the chicken α-smooth muscle actin gene is interrupted by eight introns (Carroll *et al.*, 1986). Examination of the intron positions in vertebrate α-cardiac (Hamada *et al.*, 1982; Chang *et al.*, 1985; Eldridge *et al.*, 1985), α-skeletal (Fornwald *et al.*, 1982; Zakut *et al.*, 1982; Hu *et al.*, 1986) and cytoplasmic (Bergsma *et al.*, 1985; Kost *et al.*, 1983; Nudel *et al.*, 1983; Ng *et al.*, 1985) actin genes as well as those found in sea urchin genes (Cooper & Crain, 1982; Foran *et al.*, 1985), revealed that the intron positions in these genes represent subsets of the intron positions found in the chicken α-smooth muscle actin gene (Carroll *et al.*, 1986; Table 1.2). This demonstration of an actin gene which contains all of the intron positions found in three other

Table1.2 Comparison of the intron position of deuterostome actin genes

Actin gene	Organism	Intron position							
		5'UTR	41/42	84/85	121/122	150	204	267	327/328
α-smooth	chicken[1]	X	X	X	X	X	X	X	X
α-smooth	human[2]	?	X	X	X	X	X	X	X
α-skeletal	(mouse[3] chicken[4], rat[5])	X	X			X	X	X	X
α-cardiac	chicken[6]	X	X			X	X	X	X
α-cardiac	human[7]	?	X			X	X	X	X
β-cytoplasmic	(rat[8] chicken[9], human[10])	X	X		X			X	X
SpG28	sea urchin[11]		X		X		X	X	
SpG17	sea urchin[11]				X		X		
SfA	sea urchin[12]				X		X		

Key to references :

1) Carroll *et al.*, (1986)
2) Ueyama *et al.*, (1984)
3) Hu *et al.*, (1986)
4) Fornwald *et al.*, (1982)
5) Zakut *et al.*, (1982)
6) Chang *et al.*, (1985);
7) Hamada *et al.*, (1982)
8) Nudel *et al.*, (1983)
9) Kost *et al.*, (1983)
10) Ng *et al.*, (1985)
11) Cooper *et al.*, (1982)
12) Foran *et al.*, (1985)

distinct deusterostome actin gene lineages (vertebrate striated muscle, vertebrate cytoplasmic and echinoderm) is most consistent with a scheme involving the loss of introns from common ancestral sites. It was therefore concluded, at least for the case of the deuterosome actin genes, that intron deletion has been the dominant process influencing the placement of introns in modern actin genes (Zakut *et al.*, 1982; Blake, 1983; Carroll *et al.*, 1986).

1.2 Pseudogenes

Several years ago Jacq and coworkers (Jacq *et al.*, 1977) reported the isolation and nucleotide sequence of a 5S rRNA-related gene from *Xenopus laevis* that was truncated and had mismatches when compared to the functional 5S rRNA. Jacq *et al.*, (1977) used the term *pseudogene* to describe this truncated 5S rRNA homologue. Since then many different pseudogenes have been reported from a variety of gene families, and the term can now be clearly defined as sequences found to be both related and defective (Vanin *et al.*, 1985). The varied pseudogenes reported fall into two general categories. In the first there are duplicative pseudogenes, those which are closely linked to their functional counterparts and where appropriate retain the intervening sequences of the active gene. The globin pseudogenes from a number of species form the major group within this category (Vanin, 1983). In the second and more abundant category are those lacking the intervening sequences found in their functional counterparts. Such pseudogenes have been termed processed pseudogenes for the reasons discussed below.

1.2.1 Duplicative pseudogenes

(a) The *X. laevis* 5S rRNA pseudogene

As discussed above the first gene-like sequence to be termed a pseudogene was that of the 5S rRNA described by Jacq *et al.*, (1977). The pseudogene occurs downstream of the functional 5S rRNA gene and is part of the 700 nucleotide repeat unit that is amplified during oögenesis. The pseudogene is 20 nucleotides shorter at its 3' end than its functional counterpart (101 instead of 121 nucleotides) and differs by only 9 base changes (Miller et al., 1978). No RNA corresponding to this pseudogene could be found in vivo, and thus it appeared to be an inert component of the genome. This raised the questions as to why this pseudogene structure had been conserved; whether it served some function in processing the mature 5S RNA or whether, being part of the duplicated repeat unit, it was just passively preserved along with the active gene. These questions remain largely unanswered, but the question of why no pseudogene transcripts are found in vivo has been addressed in further experiments involving microinjection of an isolated 5S gene and pseudogene into *Xenopus* oocytes. When the pseudogene is injected alone, it supports a rate of transcription of up to 85% of the level of normal 5S gene transcription. However, at least 75% of the pseudogene transcripts do not terminate correctly at the end of the gene (even although it contains a TTTT sequence thought to be important for correct termination), but read through into the adjacent sequences. In vivo this would give rise to random termination in the downstream AT-rich spacer region, and hence no discretely sized transcripts would be formed; in addition, such randomly terminated transcripts might be somewhat unstable. Thus, the lack of pseudogene transcripts of defined length *in vivo*

may be a reflection of the inefficient transcriptional termination rather than a lack of transcriptional activity per se. However a further experiment (Miller & Melton, 1981) suggests that this may not be the whole explanation. If the 5S gene and pseudogene are injected together, the rate of transcription from the pseudogene drops to one third of its level when injected alone. This indicates that there is competition between the two promoters for RNA polymerase (or other transcription factors) and the 5S gene has the more effective promoter. The two promoters only differ by four base changes and it is not clear whether this alone accounts for their different activities or whether some other feature of the enviroment surrounding the two sequences is also important.

(b) Globin pseudogenes

Historically the next set of pseudogenes to be discovered were those within the α- and β-globin gene families of different mammals (Proudfoot, 1980; Little, 1982; Lauer *et al.*, 1980; Proudfoot & Maniatis, 1980; Proudfoot *et al.*, 1982; Lacy & Maniatis, 1980; Clearly *et al.*, 1980; Clearly *et al.*, 1981; Jahn *et al.*, 1980; Fritsch *et al.*, 1980; Jeffreys *et al.*, 1982). Together the mammalian globin gene families provide examples both of pseudogenes at different stages of evolutionary decay and of the variety of processes whereby different gene clusters have evolved.

With the exception of two mouse α-globin pseudogenes that are dispersed to different chromosomes from the major α-globin gene cluster (Vanin *et al.*, 1980; Nishioka *et al.*, 1980; Leder *et al.*, 1981; Popp *et al.*, 1981), all the globin pseudogenes are found linked to their functional counterparts. The most straightforward explanation for the origin of these

pseudogenes is that they derive from duplicated genes formed within the gene clusters, which have subsequently diverged and become inactive, (i.e., transcriptionally silent). Following inactivation, such genes would have been released from selection and would then rapidly accumulate mutations at a rate more characteristic of non-coding sequences.

Estimates of the evolutionary time spent by each present day pseudogene, first under selection as an active gene and then without selection as a pseudogene, have been calculated from the percentage of silent and replacement changes in the 'coding' sequence of the pseudogene compared to the active gene (Proudfoot & Maniatis, 1980; Lacy & Maniatis, 1980; Perler *et al.*, 1980). These estimates assume that following inactivation, pseudogenes accumulate mutations at the same rate as do silent positions in active genes. However it appears that there is some selective pressure against changes, even between synonymous codons in functional genes, and that the rate of nucleotide substitution in pseudogenes is approximately twice the rate of substitutions in the third codon position of active genes (Miyata & Yasunaga, 1981; Miyata & Hayashida, 1981; Li *et al.*, 1981). Many earlier estimates do not take this factor into account and thus will have tended to overestimate the age of the pseudogene.

A further factor that has confounded these estimates is the realization that gene conversion events have played an important role in the evolution of globin gene clusters (Lauer *et al.*, 1980; Slightom *et al.*, 1980; Shen *et al.*, 1980; Leibhaber *et al.*, 1981; Schon *et al.*, 1982; Weaver *et al.*, 1981). Gene conversion is the nonreciprocal copying of information from one gene to another homologous gene within a cluster, as the result of inter- (Lauer *et al.*, 1980) or intra-chromosomal (Slightom *et al.*, 1980) exchange. A number of instances of gene conversion have been detected among α- and β-globin genes, and its effect has been to mask the true evolutionary age of genes or

pseudogenes that have undergone this conversion. Thus, two genes will appear to have arisen by duplication at the time of a conversion event, when in fact they may have a considerably older evolutionary history. For example, comparison of the proteins of two adult globins, δ and β, suggests that they arose from a duplication event not more than 40 million years (MY) ago (Spritz *et al.*, 1980; Efstratiadis *et al.*, 1980). However various non-coding regions, the second intervening sequence, the 3' untranslated region of the mRNA and 5' sequences upstream of the CCAAT box, appear to have diverged over a much longer period of time; (Martin *et al.*, 1983; Hardies *et al.*, 1984). In addition, δ-like genes or pseudogenes are found in lower primates that diverged around 75 MY ago (Jeffreys *et al.*, 1982). Thus the globin coding region appears to have undergone a recent conversion by the β-gene, which has covered the traces of its more ancient origin. Reliable estimates of evolutionary divergence times can, therefore, only be derived from those regions of the gene that have not been subjected to gene conversion.

In addition to the active embryonic (ζ) and adult (α1, α2) genes, the human α-globin gene cluster contains two pseudogenes ψζ and ψα (Lauer *et al* ., 1980; Proudfoot & Manaitis, 1980; Proudfoot *et al.*, 1982). Together, ψζ and ψα represent two extremes in the process of pseudogene formation and decay. Pseudogene ψζ shares greater than 99.5% homology in its coding region with the functional ζ-globin gene and has a single deleterious mutation, a termination codon in its first exon (Proudfoot *et al.*, 1982). Presumably it has only very recently become a pseudogene. In contrast , ψα is only 75 to 80% homologous to the active α-globin genes and has a

considerable array of mutations. These include base substitutions that introduce many missense codons and that affect the translation initiation codon, an RNA splice site and termination codons in the coding sequence, and altered spacing between CCAAT and TATA boxes in the transcriptional promoter region (Proudfoot & Maniatis, 1980). Thus ψα appears to be a relatively old pseudogene.

The human α-globin cluster also provides insight into the evolutionary mechanisms that can give rise to pseudogenes. A comparison of the sequences surrounding the ψα pseudogene and the two active genes α1 and α2 suggest that they arose by gene duplication and have subsequently undergone unequal crossing over (Lauer *et al.*, 1980; Proudfoot & Maniatis, 1980). Such events still appear to be operating in present day human populations, since chromosomes carrying either a single active α-globin gene (associated with α-thalassemia; Proudfoot, unpublished results) or an α-globin gene triplication (Higgins *et al.*, 1980; Goosens *et al.*, 1980), have been reported. Since the time that the ψα, α1, α2 cluster was formed, the two active genes α1 and α2 have been maintained closely homologous by gene conversion events, while ψα has accumulated base changes to become a pseudogene. Sequences in the intergenic regions upstream of α1 and α2 show strong homology and have been implicated in gene conversion (Proudfoot & Maniatis, 1980), and their absence upstream of ψα may perhaps explain why it too has not been subjected to conversion. Thus gene duplication by itself may not be sufficient to set a gene on the path to becoming a pseudogene; a more crucial step may be the point at which a gene no longer becomes subject to conversion by neighbouring genes and

is free to diverge on its own.

The β-globin gene clusters of a number of mammals show considerable variation in their complexity and organisation. However using the DNA sequence information available for a large majority of the β-globin genes within of species, it has been possible to relate the different present day clusters back to a simple four (or five) gene cluster , which has evolved by various gene duplication and unequal crossing-over events (Hardies *et al.*, 1984; Hardison, 1984; Goodman *et al.*, 1984).

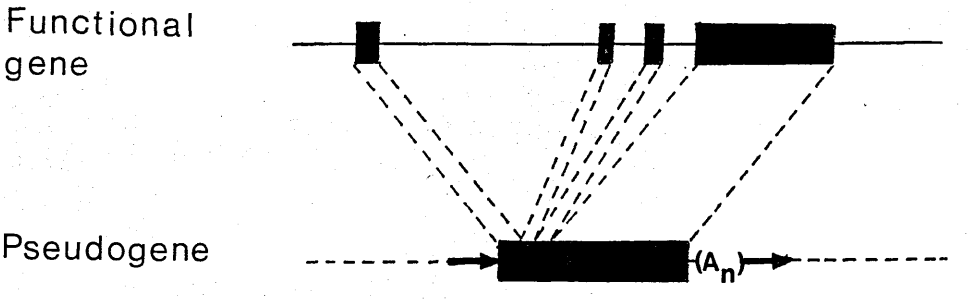
1.2.2 Processed pseudogenes

Processed pseudogenes have sequence characteristics that suggest that they were derived from the incorporation of information contained in RNA transcripts into new chromosomal locations in the genome. Processed pseudogenes relate to genes encoding proteins, but lack the intervening sequences found in the functional parent gene. Most have oligo A tracts correctly positioned relative to the poly A addition signal at their 3' ends - a feature that further points to their origin from mRNA. In addition the processed pseudogenes are found to be dispersed to chromosomal locations which generally differ from those of their parent genes. A schematic representation of a processed pseudogene and its functional counterpart is shown in Figure 1.1.

(a) Structure

Processed pseudogenes may be regarded as falling into two categories. Members of the first category are colinear with normal cellular mRNAs,

Figure 1.1 Schematic representation of a processed pseudogene and its functional counterpart



The human β-tubulin functional gene and 21β pseudogene (Gwo-Shu Lee *et al.*, 1983) are used as an example. The solid blocks represents exons, with the diagonal dashed lines between the functional gene and the pseudogene indicating the common sequences. The arrows flanking the pseudogene indicate the direct repeats. The solid line (functional gene) represents flanking and intervening sequences, while the dashed line (pseudogene) indicates the sequences flanking the pseudogene are not the same as those flanking the functional gene.

starting at the 5' mRNA cap site and ending in an A-rich or oligo A stretch of 7 to 36 nucleotides and are flanked by direct repeat sequences of 9 to 25 bases. The first example of this type was a human β-tubulin pseudogene (Wilde *et al.*, 1982a). Subsequently, similar processed pseudogenes have been found corresponding to an ever increasing number of mammalian gene families. These include pseudogenes corresponding to genes for the mouse cytochrome c (Limbach & Wu, 1985), p53 cellular tumor antigen (Benchimol *et al.*, 1984; Zakut-Houri *et al.*, 1983), and ribosomal proteins L7 (Klein & Meyuhas, 1984), L18 (Peled-Yalif *et al.*, 1984), L30 (Wiedemann & Perry, 1984) and L32 (Dudov & Perry, 1984); the rat α-tubulin (Lemischka & Sharp, 1982) and cytochrome c (Scarpulla, 1984; Scarpulla & Wu, 1983) and human β-tubulin (Gwo-Shu Lee *et al.*, 1983; Pichauntes *et al.*, 1982; Wilde *et al.*, 1982a and b), γ-actin (Leube & Gallwitz, 1986); β-actin (Moos & Gallwitz, 1982; Moos & Gallwitz, 1983), dihydrofolate reductase (Chen *et al.*, 1982; Masters *et al.*, 1983; Shimada *et al.*, 1984), arginino-succinate synthetase (Freytag *et al.*, 1984), glyceraldehyde-3-phosphate dehydrogenase (Benham *et al.*, 1984; Hanauer & Mandel, 1984), metallothionein (Karin & Richards, 1982; Varshney & Gedamu, 1984), and c-ras oncogene families (Chang *et al.*, 1982; McGrath *et al.*, 1983; Miyoshi *et al.*, 1984; Zabarovsky *et al.*, 1984). Furthermore the rat cytochrome c (Scarpulla & Wu, 1983) and human β-tubulin (Lee *et al.*, 1983) pseudogenes demonstrate that where different mRNAs with 3' untranslated regions of varying lengths are produced due to the use of alternative polyadenylation sites, processed pseudogenes corresponding to each of the different sized mRNAs may be found.

Another category of processed pseudogenes includes pseudogenes that are also clearly derived from RNA molecules, since they lack intervening

sequences found in the parent genes and end in oligo A or A-rich tracts; but with structures that do not correspond to the normal cellular mRNAs of the parent genes. There are several examples of this type: (1) a human immunoglobulin λ light chain pseudogene (Hollis *et al.*, 1983), containing spliced J and C regions but with no V region (which in immunoglobulin-producing cells is normally joined directly to the J region at the DNA level) (2) a human immunoglobulin ε heavy chain pseudogene (Ueda *et al.*, 1982; Battey *et al.*, 1982) comprising only the four spliced exons of the ε constant region but no variable region coding elements (V, D, or J regions); (3) a mouse myosin light chain pseudogene (Robert *et al.*, 1984), consisting of the five terminal exons common to both myosin alkali light chains LC1 and LC3, and lacking either of the two combinations of N terminal exons normally present in the corresponding cellular mRNA; (4) a mouse pro-opiomelanocortin pseudogene (Uhler *et al.*, 1983; Notake *et al.*, 1983) that includes only the sequences downstream of codon 67 in the 3' exon of this gene; (5) a mouse γ-actin pseudogene (Leader *et al.*, 1985), that includes only the sequence downstream from amino acid at position 7, of the actin coding region; (6) mouse cellular tumor antigen p53, where at least 80 nucleotides are missing from a long 5' untranslated reion (Benchimol *et al.*, 1984; Zakut-Houri *et al.*, 1983) and (7) mouse α-globin, α-ψ3, extends at least 350 nucleotides 5' to the transcriptional start site (Vanin *et al.*, 1980; Nishioka *et al.*, 1980).

The immunoglobulin J-C_λ and C_ε and pro-opiomelanocortin pseudogenes end in A-rich tracts of $(CA_X)_Y$ or $(GA)_X$, whereas the myosin light chain pseudogene has a short oligo A tract preceding an A-rich sequence. All are flanked by direct repeat sequences except for the mouse

α-ψ3 pseudogene. Pseudogenes (1), (2), and (3), are truncated at their 5' ends relative to their parent genes and, appear to have arisen from transcripts that initiated anomalously in the intervening sequence immediately upstream of those exons found in the pseudogene. The mouse α-ψ3 pseudogene also appears to be derived from an aberrant transcript, derived from a promoter upstream of the usual transcriptional start position.

(b) Origins

Since processed pseudogenes are found in all, or most, individuals of a species and are transmitted as inheritable components of the genome, they must have originally arisen in cells of the germ line. It follows from this that processed pseudogenes would be expected only to be formed from those genes which are expressed in the germ line cells. Indeed, in the main, those processed pseudogenes that are essentially colinear with cellular mRNAs do seem to be derived from either 'housekeeping' genes common to all cell types or from genes that might be preferentially expressed in the germ line (e.g., tumour antigen p53, *c-ras* oncogenes).

In contrast the majority of processed pseudogenes that appear to derive from aberrant transcripts, originate from genes that are not normally expressed in the germ line since they encode products of highly differentiated somatic cells (i.e., lymphocyte immunoglobulin chains, erythrocyte α-globin). Presumably the aberrant nature of the transcripts from which they appear to be derived is a reflection of their abnormal transcription in the germ line.

The human and mouse actin genes further exemplify very clearly this point that processed pseudogenes are usually only found in gene

families that are expressed in the germ line. Processed pseudogenes appear to account for a large part of the genomic sequences related to cytoskeletal β- and γ-actins, which are expressed in all non-muscle cells (Moos & Gallwitz, 1982; Moos & Gallwitz, 1983; Ponte *et al.*, 1983; Minty *et al.*, 1983). In contrast, there are no pseudogenes corresponding to the α-cardiac and α-skeletal muscle actins, products of differentiated somatic tissues (Ponte *et al.*, 1983). There are several examples, including those of mouse ribosomal proteins L7, L18, and L32, (Klein & Meynhas, 1984; Perled-Yalif *et al.*, 1984; Dudov & Perry, 1984), human non-muscle tropomyosins (MacLeod & Talbot, 1983), a β-tubulin isotype (Lee *et al.*, 1983), and arginosuccinate synthetase (Freytag *et al.*, 1984), comprising a single active gene and anything from 3 to 15 processed pseudogene counterparts. The number of pseudogenes corresponding to any one protein may be a reflection of the relative extent of transcription of the functional gene in the germ line (Lee *et al.*, 1983).

Almost all processed pseudogenes have been found in mammalian species. However a single calmodulin processed gene has been found in chickens (Stein *et al.*, 1983), and one, at least, of the histone 'orphons' of sea urchins is derived from reverse transcribed mRNA (Liebermann *et al.*, 1983). In addition the F elements of *Drosophila melanogaster* appear to be dispersed by the integration of polyadenylated RNA transcripts (DiNocera *et al.*, 1983). Therefore the mechanisms responsible for the generation of processed pseudogenes are not exclusive to mammals, although some features of mammalian gamete production and germ line transcription may make them peculiarly susceptible to the formation of processed pseudogenes.

(c) Age and divergence

Unlike duplicative pseudogenes, which may be as little as 75% homologous to the parent genes, processed pseudogenes seem to show strikingly high (90 to 99%) homology to the genes from which they were derived. This suggests that they have arisen relatively recently in evolutionary history.

The myosin light chain pseudogene, for example, shares 99% nucleotide sequence homology with the active gene and, furthermore, is found in *Mus musculus*, but not the related species *Mus spretus*, which diverged less than 7 MY ago (Robert *et al.*, 1984). Similarly, a set of three human β-tubulin pseudogenes show homologies of 91, 92 and 97% with their parent gene, and it has been estimated that they diverged around 13.4, 10.7 and 4.4 MY ago, respectively (Lee *et al.*, 1983). A further indication of the relative recent origin of some processed pseudogenes is the observation that a human dihydrofolate reductase pseudogene, hDHFR-ψ1, which has perfect homology to the functional gene, is only present in certain individuals of the species and shows an imbalance in its frequency in different racial groups (Anagnou *et al.*, 1984).

Thus processed pseudogenes appear to be recent genomic acquisitions. However, because the examples of processed pseudogenes studied to date have been detected and isolated using DNA hybridisation probes, the sample may be somewhat biased towards those that are little diverged from their parent genes. If probes were used at high stringency, more diverged processed pseudogenes may well have gone unnoticed. Indeed, when genomic blots are performed at reduced stringency, additional genomic sequences with weaker homology to a probe can often be seen (Lee *et al.*, 1983; Minty *et al.*, 1983). Furthermore an example of a highly divergent

processed pseudogene with only 77 to 80% nucleotide homology to an active β-tubulin gene has been isolated from a human genomic library (Wilde *et al.*, 1982). Therefore, genomes may contain whole series of processed pseudogenes that have become progressively more and more diverged from their parent genes, gradually 'fading out' into the genomic background.

(d) Expression ?

It has been assumed that processed pseudogenes will have been transcriptional inactive since their time of formation. With the exception of the mouse α-ψ3 globin pseudogene, which retains upstream RNA polymerase II promoter sequences, all other processed pseudogenes are coterminal with their corresponding mRNAs and thus lack transcriptional promoters. Although it is not impossible to envisage integration occurring correctly downstream of an RNA polymerase II promoter, it seems unlikely that this could occur without adversely affecting the activity of other genes. Thus, it is simplest to assume that, almost by definition, processed pseudogenes will have been incapable of expression from the time of their formation onwards, even though initially they will have had intact coding regions and only subsequently acquired the deleterious mutations characteristic of 'duplicative' pseudogenes. Consistent with their inertness, some pseudogenes show a higher degree of DNA methylation than their functional counterparts (Lund & Dahlberg, 1984; Dudov & Perry, 1984).

In view of these considerations, it is somewhat surprising that a processed calmodulin 'pseudogene' appears to be specifically expressed in chicken muscle (Stein *et al.*, 1983). However, clarification of this observation awaits the nucleotide sequence of regions flanking this processed pseudogene and a more detailed structural analysis of the reported

tissue specific transcript. There are however other proposed examples of functional processed pseudogenes, rat preproinsulin I gene (Soares *et al.*, 1985) and *chironomus* globin gene (Antoine & Niessing, 1984).

(e) The snRNA pseudogenes

Although only processed pseudogenes derived from genes encoding proteins have been discussed here it is interesting to note that they share structural features with snRNA pseudogenes. Small nuclear RNAs (snRNA) are a family of abundant discrete RNAs found associated with proteins in ribonucleoprotein particles in the nuclei of eukaryotes. Each snRNA species (U1, U2, U3, U4 and U6 RNAs) is apparently encoded by approximately 100 to 2000 genes that are dispersed in the genome, these estimates being based on solution hybridisation experiments and on the frequency of clones in bacteriophage genomic libraries that hybridise to snRNAs (Hayashi, 1981; Westin *et al.*, 1981, Denison *et al.*, 1981). However sequence analysis of a number of cloned fragments hybridising to the snRNAs, revealed that, the vast majority contained snRNA pseudogenes (Hayashi, 1981; Westin *et al.*, 1981, Denison *et al.*, 1981), perhaps as many as 80 to 90% of genomic sequences are pseudogenes.

The pseudogenes are of several different types, classified on the basis of their structural characteristics. Some encode full-length snRNAs, but contain scattered base substitutions and insertions (Westin *et al.*, 1981; Manser & Gesteland, 1981; Monstein *et al.*, 1983). In view of the virtual invariance of snRNAs in evolution, it appears unlikely that these sequences encode functional snRNAs. These pseudogenes also show significant homology to functional snRNAs in their flanking regions, suggesting they were generated by divergence of duplicated snRNA genes. The significantly

greater conservation of 'coding' as apposed to flanking sequences even in the pseudogenes perhaps indicates that gene conversion has also been operating in this dispersed gene family (Denison & Weiner, 1982).

Other snRNA pseudogenes, in contrast, have characteristics that led to the suggestion that they were generated by the incorporation of reverse transcripts of snRNAs into the genome at either blunt or staggered chromosomal breaks (Van Arsdell *et al.*, 1981). A number of different mechanisms for the integration process have been elaborated to take into account the different flanking structures of these pseudogenes; these are discussed more fully below. These pseudogenes are characterised by only containing sequences that are present in snRNA molecules themselves; their homology with snRNA genes begins precisely at the snRNA 5' end and extends either to the 3' end of the snRNA or shows a slight or more severe degree of 3' truncation. Some, but not all, pseudogenes are flanked by short direct repeats of 16 to 21 nucleotides; the longest snRNA pseudogenes additionally have short 3' A-rich segments at their ends or preceding a 3' direct repeat sequence (Hayashi, 1981; Piechaczyk *et al.*, 1982). Since poly A is not normally present on snRNAs, such pseudogenes must have been derived from aberrantly polyadenylated molecules.

(f) Mechanism of insertion

The basic mechanism whereby processed pseudogenes are formed has been taken as the insertion of an mRNA or its cDNA copy into a staggered (or blunt) break in chromosomal DNA and subsequent repair of single stranded regions. While this outlined mechanism has gained wide acceptance, it has been considerably more difficult to define in greater detail the precise series of molecular events that give rise to these pseudogenes, since the only

information concerning their mechanism of origin derives from the organisation of sequences flanking them.

Any model for the formation of these pseudogenes must address the following questions: What is the polymerase responsible for the reverse transcription ? How is the reaction primed ? Where and how do the insertions occur in the genome ? Is the inserted molecule an RNA or a cDNA (or an RNA-cDNA heteroduplex) ?

The reverse transcriptase activity responsible for the formation of these RNA - derived pseudogenes could have come from an endogenous retrovirus or a transient germ line infection by a retrovirus (Berstein *et al.*, 1983). It seems equally possible that they are formed as the result of some secondary activity of normal cellular DNA polymerase since human DNA polymerase β can copy synthetic RNA template in vitro (Weissbach, 1977). However a source of cellular reverse transcriptase activity may be provided by the long interspersed repeated sequences, (L1 elements) which have recently been reported to have the potential to encode a protein with such activity (Loeb *et al.*, 1986).

The sites into which processed pseudogenes have integrated are often found to comprise relatively AT-rich sequences, as indicated by the direct repeat flanking sequences. Examples of such repeat sequences of processed peudogenes are shown in Table 1.3. Since such sequences are more prone to local melting of DNA strands and hence strand breakage, they might be expected to be a common source of sites for pseudogene insertion. It has also been suggested that topoisomerases play an important role in generating transient breaks in DNA between which the insertion may occur (Van Arsdell & Weiner, 1984).

Questions concerning the primer for reverse transcription and the nature of the inserted molecule will be discussed together in comparing

different models (shown in Figure 1.2), proposed to account for pseudogene formation. The first model (Figure 1.2A), that of Van Arsdell *et al.* for snRNA pseudogenes (Van Arsdell *et al.*, 1981), suggested the following sequence of events (1) synthesis of a cDNA copy of the snRNA; (2) covalent linkage of the 3' end of the cDNA to a 5' overhang of a staggered chromosomal break; (3) second strand cDNA synthesis primed from the recessed 3' OH of the break; and (4) ligation and repair of the ends of the break, creating flanking direct repeats. The authors preferred the insertion of a reverse transcript of the snRNA molecule as this obviated the need to propose mechanisms for decapping the snRNA and for the ligation of RNA to DNA. Of itself this model does not explain how synthesis of the first cDNA strand is primed. For severely truncated snRNA this presents no problem since the snRNAs from which they derive can act as self-priming templates for reverse transcriptase *in vitro* (Berstein *et al.*, 1983); and if similar cDNAs were formed *in vivo*, they could give rise to pseudogenes as indicated in the model. However in extending this model to full-length snRNA pseudogenes and to processed pseudogenes that are full-length copies of mRNAs, it is presumably necessary to invoke some exogenous T-rich primer molecule for synthesis of the first cDNA strand.

This minimal 'cDNA insertion' model has been elaborated to involve topoisomerases in the formation of staggered or blunt chromosomal breaks (Van Arsdell *et al.*, 1981; Van Arsdell & Weiner, 1984; Figure 1.2B). In addition, it was suggested that homology between the downstream direct repeat sequence and the incoming cDNA molecule might be instrumental in anchoring the cDNA relative to the staggered break (Moos & Gallwitz, 1983). This would account for the fact that flanking direct repeat sequences frequently overlap the 3' end of truncated U2 snRNA pseudogenes or the 3' oligo A or A-rich tails of full-length snRNA and processed pseudogenes,

Key to references :

1) Gwo-Shu Lee *et al.*, (1983)

2) Wilde *et al.*, (1982)

3) Moos & Gallwitz, (1982)

4) Moos & Gallwitz, (1983)

5) Hollis *et al.*, (1982)

6) Battey *et al.*, (1982)

7) Varshney & Gedamu, (1984)

8) Lemischka & Sharp, (1982)

9) Scarpulla, (1984)

10) Zakut-Houri *et al.*, (1983)

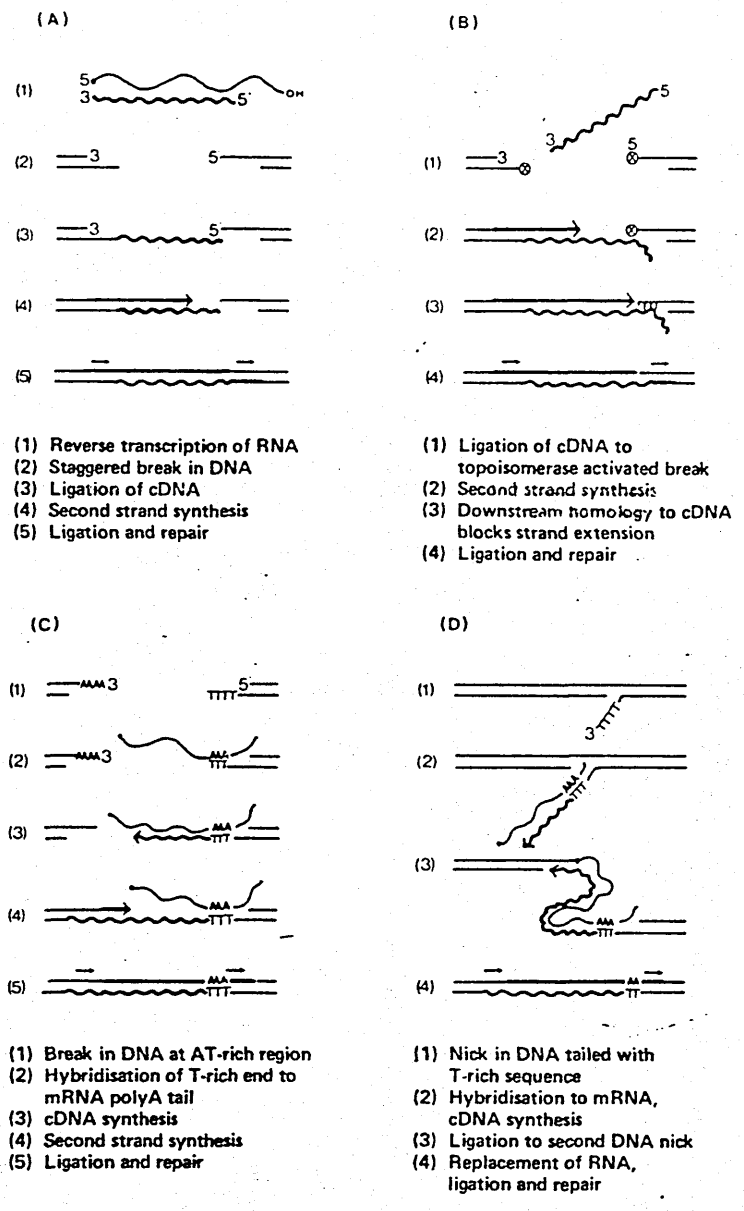
11) Wiedemann & Perry, (1984)

Table 1.3 Sequence of direct repeats flanking processed pseudogenes

Processed pseudogene	Flanking direct repeats 5' repeat	3' repeat
Human		
7β-tubulin[1]	CAATAAAATGCACAGGTCTGCC	AAAAAAAATGCACAGTTCTACA
11β-tubulin[2]	CACTCAAAGAAATCAGAGATGT	AAAAAAAAGAAATCAGAGACTG
ψ1 β-actin[3]	CATATAAAACTTATGTTTCTGC	AAAAAAACACTTATGTTTCCAC
ψ2 β-actin[4]	ATATATAAACCTCCTTACACCG	AAAAAAAAACCTCCTTGCATAT
λψ1 immunoglobulin[5]	CTTAGAAGAGGATGTGAATGCT	AAAAAAAGAAGATGTGAATATT
ε immunoglobulin[6]	CAAATTGTGCCTAAGCGAATTT	ACACTAAAACCTAGAGGAAAAC
methallothionein I[7]	TTTAAAGAGGTAATTAAGGCAC	AAAAAAAAGGTAATGAAGGGTG
Rat		
α-tubulin[8]	CTTATAAAAAGAGATTTTTGGC	CTTAAAAAAAAGAGATTTTTTTT
RC-5 cytochrome c[9]	GAGCTCATAAAGACCTGTAGCC	ATTTAAAAAAAGAACTGTAACC
Mouse		
p53 tumour antigen[10]	CTCTATAAAGAACTCAAGAGGT	AAAAAAAAAGAACTCAAGAAAC
ribosomal protein L30[11]	AATGAAAACTCTAACATTCGCC	AAACAAAACTCTAACATTCTCC
ribosomal protein L32[11]	ACATTACAAATTAGCTGCTGCT	AAAAAACAAATTAGCTGCTTTT

* The direct repeats are overlined.

Figure1.2 Models proposed for the generation of RNA-derived processed pseudogenes.



Thin wavy lines represent RNA, thick wavy lines cDNA, and thick lines second strand or repair DNA synthesis. Flanking direct repeats resulting from the insertion are indicated by short arrows (---->) and topoisomerase molecules by ⊗ . (A), (B) 'cDNA insertion' models for the generation of snRNA pseudogenes (Van Arsdell *et al.*, 1981; Van Arsdell & Weiner, 1984), (C) 'Primed insertion' model for mRNA derived pseudogenes (Vanin, 1984). (D) Retroposon insertion (Rogers, 1985).

(Table 1.3).

This latter observation also points to an alternative model, which to a large extent overcomes the difficulty of 'cDNA' insertion. The overlap between the 3' ends of pseudogenes and their flanking direct repeats, suggests that the 3' overhangs at staggered chromosomal breaks might themselves act as primers for the initial cDNA synthesis by virtue of their partial homology to RNA. Thus this model (Figure 1.2C), combines the two steps of cDNA synthesis and cDNA insertion. Since the cDNA molecule is primed by a single stranded region of the genomic DNA itself, it is necessarily already linked into the chromosome. Subsequent steps would involve the replacement of the RNA to generate a double-stranded cDNA and repair and ligation of the ends (Vanin, 1984).

A variation on this 'primed insertion' theme has been suggested by Rogers in a general model for retroposon formation (Rogers, 1985). In this model, (Figure 1.2D), a nick in chromosomal DNA becomes tailed with T-rich sequences, which then act as primers for cDNA synthesis. To ensure complete copy of the mRNA, the 5' end of the inserted RNA is ligated to a second nick in the target DNA and repair synthesis completes the process to generate a retroposon flanked by direct repeats.

It is most likely that no one mechansim is universal, and the variety of pseudogenes and retroposon structures and flanking 'tail' and repeat sequences probably reflects a variety of ways in which sequences contained in RNA may be reintroduced into the genome.

1.3 Eukaryotic repetitive DNA

Prokaryotes possess relatively small genomes consisting predominately of DNA sequences of low copy number. The sizes of the genomes of different species vary by less than an order of magnitude (Kingsbury, 1969). Eukaryotic genomes are generally much larger than their prokaryotic counterparts, and a far greater proportion (30-40%) of their DNA is repeated (Britten & Kohne, 1968; Laird, 1971). This repetitive component consists of several types of sequence and it has often been useful to classify these sequences according to their structure, distribution and frequency of repetition (Jelinek & Schmid, 1982).

The repetitive sequences of the eukaryotic genome can be divided into highly repetitive and middle repetitive fractions on the basis of renaturation kinetics (Britten & Kohne, 1968). The highly repetitive fraction consists of what are termed DNA satellites, generally the most highly repetitive sequence component of the eukaryotic genome (Britten & Kohne, 1968). 'Middle-repetitive' DNA is a term used as a broad description of heterogenous sequence components consisting of many different families of lower copy number repetitive DNA (Britten & Kohne, 1968).

Although the main purpose of this section is to review 'middle-repetitive' DNA sequences, it is also appropriate to present a brief summary of the main features of satellite DNA.

1.3.1 DNA satellites

Satellite DNA represents highly repeated sequences, of which there may be a million or more copies per haploid genome, which are usually quite short and are arranged in tandem arrays. The origin of the name satellite relates to the method of its isolation on caesium chloride buoyant density gradients of sheared DNA where it will sometimes form a satellite band separated from the main DNA band, due to its differing content of adenine and thymine residues. The simplest known satellite DNA is poly [d(A-T)] which occurs in certain crabs. Other satellites can have any number up to several hundred base pairs which are repeated in tandem fashion along the genome.

Human DNA has been shown by density-gradient centrifugation to have four main satellites (Jones & Corneo, 1971; Evans *et al.*, 1974), and by dye binding (Ohno, 1971) and restriction endonuclease cleavage (Maio *et al.*, 1977), to have two additional satellites. The distribution of the satellite DNA among chromosomes varies. Some chromosomes have virtually no satellite sequences while others (notably the Y chromosome) are largely composed of satellite sequences (Miklos & John, 1979). In general satellite DNA appears to be concentrated near the centromere of the chromosomes in the heterochromatin fraction. DNA sequence analysis has shown that the basic repeat unit of satellites is itself made up of subrepeats. For example the major mouse satellite has a repeating structure of 234 base pairs made up of four related 58 and 60bp segments each in turn made up of 28 and 30bp sequences (Manuelidis, 1978; Horz & Altenburger, 1981). The satellites between and within related species are themselves related in an evolutionary sense by cyclical rounds of multiplication and divergence of

an initial short sequence (Southern, 1975). The nature of the multiplication process is not known for certain but probably involves unequal recombination events. The divergence involves single base changes and insertions and deletions (Pech *et al.*, 1979; Taparowsky & Gerbi, 1982).

Despite the detailed knowledge of the sequence and distribution of satelllite DNA, there is little idea as to the function in the cell. Originally, it was thought that satellite DNA was not transcribed since RNA of corresponding sequence was seldom isolated. However occasional cases of satellite transcription have since been reported (Varley *et al.*, 1980; Jamrich *et al.*, 1983). On the whole, transcriptional inactivity of satellite DNA ties in with its localisation in heterochromatin. As satellite DNA is often lost in somatic cells, it has been proposed that it may have some function in the germ cells (Gautier *et al.*, 1977; Adams *et al.*, 1983; Bostock, 1980). This function may relate to the recombination events which occur during gametogenesis and which may be enhanced by the presence of blocks of similar DNA sequences on several chromosomes.

1.3.2 Middle-repetitive DNA

'Middle-repetitive' DNA is a term usually used as a broad description of an additional heterogenous sequence component consisiting of many different families of lower-copy-number repetitive elements which collectively comprise a major fraction (30 - 40%) of the DNA in most eukaryotic genomes (Britten & Kohne, 1968). Middle-repetitive DNA has been studied in a variety of eukaryotes. However this review will look mainly at new studies in a few selected organisms, in which the greatest

advances in understanding the structure and distribution of middle-repetitive DNA have been made.

(a) *Drosophila* middle-repetitive DNA

Approximately 12% of the genome of *Drosophila melanogaster* consists of 'middle-repetitive' DNA (Brutlag *et al.*, 1977). About one quarter of this component consists of dispersed tRNA genes and tandemly-repeat genes coding for histones, rRNA and 5s RNA. The remainder consists of about 50 or more families of dispersed repeated elements containing between 10 and 100 sequences per family. Using a panel of seventeen dispersed middle-repetitive DNA sequences selected at random by cloning, Young (1979), showed that the location of some or all differed in the polytene chromosomes of two non-interbreeding strains of *Drosophila melanogaster*, indicating that in all cases the sequences were derived from families of mobile genetic elements. Similar experiments have been performed in several other laboratories (Rubin *et al.*, 1981; Ananiev *et al.*, 1984; Hunt *et al.*, 1984; Junakovic *et al.*, 1984). Some of these sequences corresponded to well characterised families of transposable genetic elements including *copia*-like sequences (*Copia*, 412, 297, 17.6, mdgl, mgd3, b104 (Rubin *et al.*, 1981; Scherer *et al.*, 1982) and other distinct families of mobile elements including FB elements (Potter *et al.*, 1980), Gipsy (Modolell *et al.*, 1983), P-elements (Rubin *et al.*, 1982), hobo (McGinnis *et al.*, 1983), I-factors (Bucheton *et al.*, 1984) and less well-characterised mobile elements (Young, 1979). It has been estimated that these families of dispersed transposable genetic elements collectively may total over 30 and account for most of the remaining 75% of the middle-repetitive DNA in *Drosophila melanogaster* and related species (Spradling and Rubin, 1981). The locations of these dispersed mobile

elements are generally conserved within an inbred fly population (Ananiev *et al.*, 1984) and invariant between separate stocks of the same species (Young, 1979; Junakovic *et al.*, 1984). Moreover, some families of transposable elements may be absent altogether from closely related species of *Drosophila* (Dowsett & Young, 1982; Hunt *et al.*, 1984). The remainder of the middle-repetitive elements appear to be confined to constant positions at specific chromosomal locations, including the pericentromeric regions of polytene chromosomes (Dowsett & Young, 1982). Recent careful studies (Ananiev, *et al.*,1984) have revealed several other significant findings concerning the properties of mobile dispersed middle-repetitive elements in *Drosophila*. These include the observation that some families of elements may 'prefer' to transpose into similar genomic locations; that the presence of a number of such elements at a single chromosomal region does not affect chromosome morphology; that polytene bands with the largest DNA contents probably offer the largest targets for transposition; that the regions of DNA surrounding centromeres may be composed almost entirely of clusters of mobile elements.

From this large amount of structural information it is possible to come to several conclusions; (1) the majority of middle-repetitive DNA in *Drosophila* consists of potentially mobile genetic elements; (2) the chromosomal location and copy number of a given mobile middle-repetitive element is under close genetic control within a given fly population, and (3) most of the dispersed middle-repetitive DNA provides no function essential to the survival of these insects.

(b) Rodent and primate middle-repetitive DNA

Mammalian middle-repetitive DNA can generally be classified into two categories according to the length of the repeating unit.

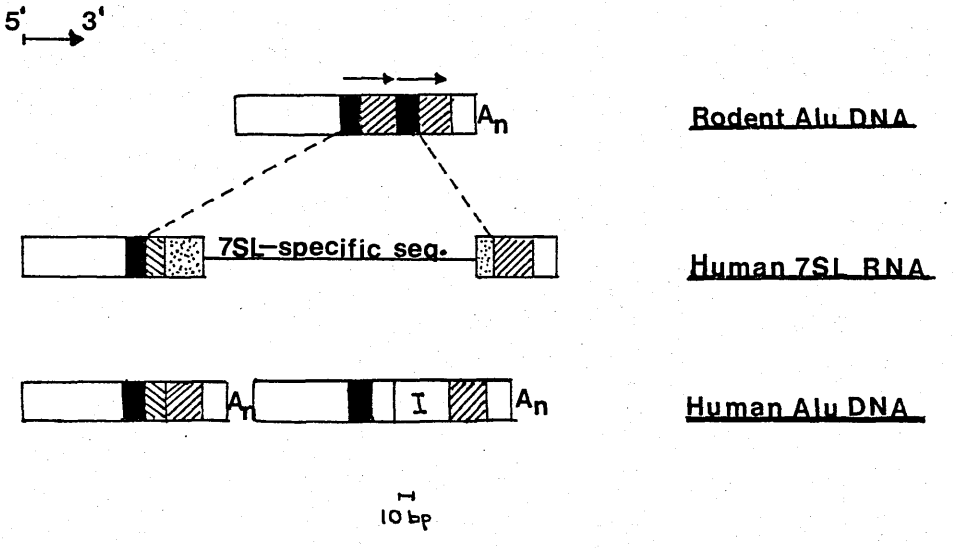
(1) SINEs, short interspersed repetitive elements that are normally several hundred base pairs in length.

(2) LINEs, long interspersed repetitive elements which appear to be thousands of base pairs in length (Singer, 1982).

This section will concentrate on the two most abundant and well characterised members of these middle-repetitive sequence families in mammalian DNA: the short interspersed AluI repeats (Houck *et al.*, 1979), and the long interspersed repeated elements referred to as LINE or L1 elements (Voliva *et al.*, 1983; Singer, 1982, Singer & Skowronski, 1985).

(i) *Alu*I-repeats. Most of the middle-repetitive DNA in mammalian genomes consists of numerous families which are only a few hundred base pairs in length (Schimd & Deininger, 1975). One SINE family dominates this repetitive fraction and is referred to as the *Alu*I family because most of its members contain *Alu*I restriction sites (Houck, et al., 1979). There are 500,000 copies of *Alu*I-repeats, representing several percent of the genome. Equivalent sequences to AluI-repeats have been identified in other primates and in rodents (Grimaldi *et al.*, 1981; Hayes *et al.*, 1981), and also *Xenopus* (Ullu & Tschudi, 1984). Human AluI-repeats consist of a head-to-tail tandem arrangement of two related sequence about 130bp long, each terminated by an A-rich tail. This is shown diagrammatically in Figure 1.3. One of the sequences contains an additional, internal segment of 32bp (Deininger *et al.*, 1981). The equivalent sequence in rodents is derived from just one 130bp repeating unit, containing a tandem repeat form by a duplication of a internal 30bp sequence (Kalb *et al.*, 1983).

Figure 1.3 The structural relationship of human 7SL RNA to the consensus sequence of human and rodent *Alu* DNA



Homologous sequences are indicated by identical shading. Human *Alu* DNA is a head to tail dimer of two similar sequences, about 130bp long. The right monomer contains an insert (I) which is not present in the left half (Deininger *et al.*, 1981). The rodent *Alu* equivalent sequence is a monomer (Krayev *et al.*, 1980; Haynes *et al.*, 1981). The mouse B1 *Alu* -equivalent consensus sequence compiled by Kalb *et al.*, (1983) contains an internal tandem duplication of 30bp. Arrows above the rodent *Alu* DNA indicate the position of the 30bp tandem duplication; $(A)_n$ denotes an A-rich sequence which follows the *Alu* sequence at the 3' end.

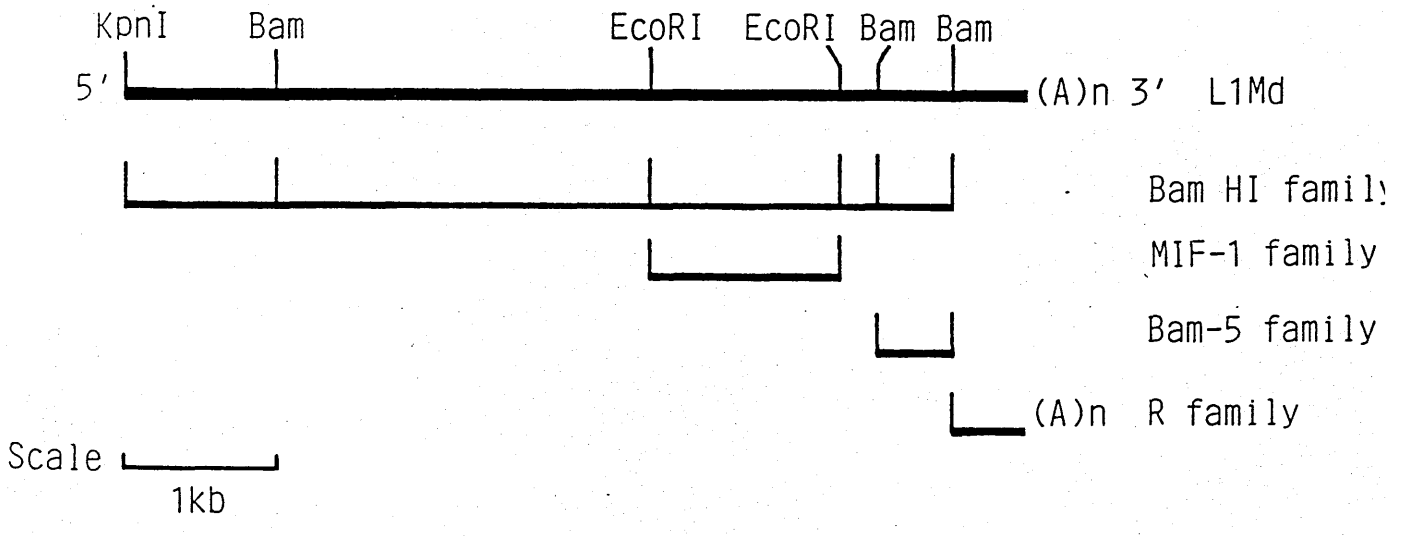
Recent studies have revealed a high sequence homology (80%) between the longer unit of the AluI consensus sequence and the 5' and 3' portions of the 7SL RNA. The 7SL RNA is an abundant cytoplasmic RNA, 300bp in length and forms part of the signal recognition particle, (Walter & Blobel, 1980). As shown in Figure 1.3 the central 155bp of the 7SL RNA sequence is absent from the AluI-repeat (Ullu & Tshchudi, 1984). This work provided an important insight into the evolution of AluI repeats in mammalian DNA. Only two 7SL RNA genes and no AluI-repeats are found in the *Drosophila* genome (Gundelfinger *et al.*, 1984). Analysis of the 7SL RNA in man, *Xenopus* and *Drosophila* indicated that the sequence is subject to strong evolutionary conservation (Ullu & Tschudi, 1984). It has been argued that AluI-repeats are derived from processed 7SL RNA transcripts, containing a 3' poly A tail (Gundelfinger, 1983). Altogether this research has suggested that 7SL RNA is the progenitor of the *Alu*I sequence family.

(ii) L1 elements (Rogers, 1984; Singer & Skowronski, 1985). Primate and rodent DNA appears to have only one major family of long interspersed repeated elements, and is referred to as the L1 family. Primate L1 sequences have shown to be evolutionary related to the L1 family of rodents by DNA hybridisation and sequence analysis (Manuelidis & Biro, 1982; Martin *et al.*,1984; Singer *et al.*, 1983). More recently it has been shown that sequences homologous to L1 elements are present in a wide variety of mammals, suggesting that L1 is ancient and has been conserved through mammalian evolution (Katzir *et al.*, 1985; Witney & Furano, 1984). There are 10^5 copies of these elements, varying in length up to 6 - 7kb and accounts for at least 2 - 3% of the mammalian genome (Singer, 1982). Different segments of rodent L1 element were cloned independently as separate sequence and were

referred to as BstN1 (Cheng & Schildkraut, 1980), BamHI (Soriano *et al.*, 1983), Bam5 (Fanning, 1982), R-family (Gebhart, Zachau, 1983) and MIF-1 repeats (Brown, 1983). These separate repeat sequences were then later shown to be colinear (Fanning, 1983; Bennett & Hastie, 1984; see Figure 1.4).

The majority of the members of the human and rodent L1 families are not 'full-length' copies of the consensus sequence (Figure 1.4). Most members are truncated at different and apparently random distance from a common 3' end (Fanning, 1983; Voliva *et al.*, 1983). Therefore, the extreme 5' sequences of the L1 elements are represented less frequently (approximately 10,000 times in the genome) than extreme 3' sequences (85,000 times in the genome), (Gebhard *et al.*, 1982). The 3' end of individual L1 elements contain a poly A tail of variable length (Lerman, 1983; Grimaldi, et al., 1984), which corresponds to the 3' end of RNA transcripts in vivo (DiGiovanni *et al.*, 1983). Individual L1 elements are bordered by small, (less than 15bp) direct repeats. Taken together these observations suggest that individual L1 elements are generated via an RNA intermediate and insert at a staggered break in the genome (Voliva, *et al.*, 1984; Wilson & Storb, 1983). Several 'full-length' mouse L1 elements have recently been isolated (Loeb *et al.*, 1986). Comparison of their 5' ends revealed that L1 elements have multiple copies of a 208bp direct tandem repeat at their 5' end. The two examples documented so far, have 4 $^{2}/_{3}$ and 1 $^{2}/_{3}$ copies respectively of the tandem repeat, the $^{2}/_{3}$ copy being the most 5' member (Loeb *et al.*, 1986). Hybridisation experiments indicate that this 208bp sequence is a regular feature of many long L1Md members. However this tandem repeat shows no homology with a previously described 5' end of a L1Md element which was also internally and genomically repetitive (Fanning, 1983). The presence of at least two different ends could be an indication of different biological functions.

Figure 1.4 Consensus restriction map of L1 elements



Several investigators have noted an open reading frame in both primate and mouse L1 (Manuelidis, 1982; Martin *et al.*, 1984; Potter, 1984). Martin *et al.* (1984) compared a 312bp region of monkey and mouse L1 sequences, finding a silent versus replacement ratio indicating that this portion of L1 has evolved under the selection for protein function. Recent anaylsis of 'full-length' mouse L1 elements has identified two large open reading frames (ORFs) of 1,137 and 3,900bp which are also evolving under the selection of protein function (Loeb *et al.*, 1986). An open reading frame homologous to the larger ORF of the mouse L1 element has also been identified in primate L1 elements (Hattori *et al.*, 1986). It was shown that the rodent and primate L1 elements have significant sequence homology to several RNA dependent DNA polymerases of viral and transposable element origin (Loeb *et al.*, 1986; Hattori *et al.*,, 1986). This provides a possible explanation for the preferential active dispersion of the L1 family sequence.

The present state of knowledge of L1 elements leaves several issues unresolved. The main issue concerns the function of the L1 element gene product. Correlating genotype and phenotype, which is difficult to do in a mammalian genetic system, is made even more difficult by the properties of the L1 family. It is difficult to isolate a functional L1 gene because of the copy number and the homogeniety of the family. Rodent L1 transcripts can identified but they appear to be heterogeneous in length (Fanning, 1982; Soriano, *et al.*, 1983) and are transcribed from both strands (Jackson *et al.*, 1985). Transcription studies of primate L1 indicate both heterogenous-sized (Kole *et al.*, 1983; Shafit-Zagardo *et al.*, 1983) and homogenous-sized (Kole *et al.*, 1983; Skowronski & Singer, 1985) strand-specific RNAs can be found. So far no L1 protein products have been identified.

(c) Foldback DNA

'Foldback DNA' is a term originally coined by Wilson & Thomas (1974) to describe the DNA structures formed when eukaryotic DNA is denatured and allowed to anneal at low DNA concentrations to avoid intermolecular reassociation. They result from the presence of inverted repeat sequences located within the same DNA fragment, and account for a variable though significant fraction (1-10%) of the DNA in most eukaryotic genomes. It is established that foldback DNA is represented in all frequency classes and is widely distributed throughout metaphase chromosomes. Its size is generally in the range 300 to 1200 base pairs, although in some cases it can be as large as several kb (Perlman *et al.*, 1976; Jelinek, 1978; Schmid & Deininger, 1975; Hardman *et al.*, 1979a,b). Initially the general properties and distribution of foldback sequences were studied in a wide range of eukaryotes, from slime moulds to mammals (Cech & Hearst, 1975; Deininger & Schmid, 1975; Hardman & Jack, 1977). In this early work much attention was paid to studying differences in the distribution of foldback elements in different species and making correlations between the properties of foldback DNA and middle-repetitive DNA sequences (Schmid *et al.*, 1975; Hardman *et al.*, 1979b, 1980). Just over a decade ago a reassociation-kinetic study of total *Xenopus laevis* foldback DNA led to the suggestion that these sequences may be mobile genetic elements (Perlman *et al.*, 1976).

1.4 Background and objectives of this research project

The objectives of the work described in this thesis were to analyse two examples of repeated DNA. They were an example of inverted repeat DNA and a possible large repeated DNA region, both of which were in mouse genomic clones containing actin-like sequences, presumed to represent processed pseudogenes.

The initial observations which provoked this work and provided one of its bases were obtained by electron microscopic heteroduplex analysis of the two clones isolated from a mouse genomic lambda library by screening with an actin cDNA probe. The analysis was performed by Dr H.Delius (EMBL Heidelberg) and was initially undertaken to locate the actin-like regions within the genomic clones λmA14 and λmA36. Individual separated DNA strands from the lambda recombinants were annealed to one of two reference mouse genomic clones, λmA19 and λmA81, which were known to contain γ-actin processed pseudogenes in different orientations relative to the lambda arms. The positions of these actin pseudogenes within the mouse DNA inserts were known, and the complete sequence of the actin pseudogene in λmA19 was subsequently determined (Leader *et al.*, 1985). Thus, measurements of the position of the heteroduplex formed between the actin-like sequences in λmA14 and λmA36, and the reference pseudogene with the same orientation, allowed the position of the actin-like DNA relative to the lambda arms to be deduced from λmA14 and λmA36.

Figure 1.5 shows an electron micrograph of a heteroduplex between separated single strands of λmA19 and λmA14, and Figure 1.6 is a schematic interpretation of this. It can be seen that the actin-like regions in the two

Figure 1.5 Electron micrograph of the heteroduplex formed between separated single strands of λmA19 and λmA14

The electron micrograph is courtesy of Dr H.Delius (EMBL Heidelberg). A schematic interpretation of this micrograph is shown in Figure 1.6.

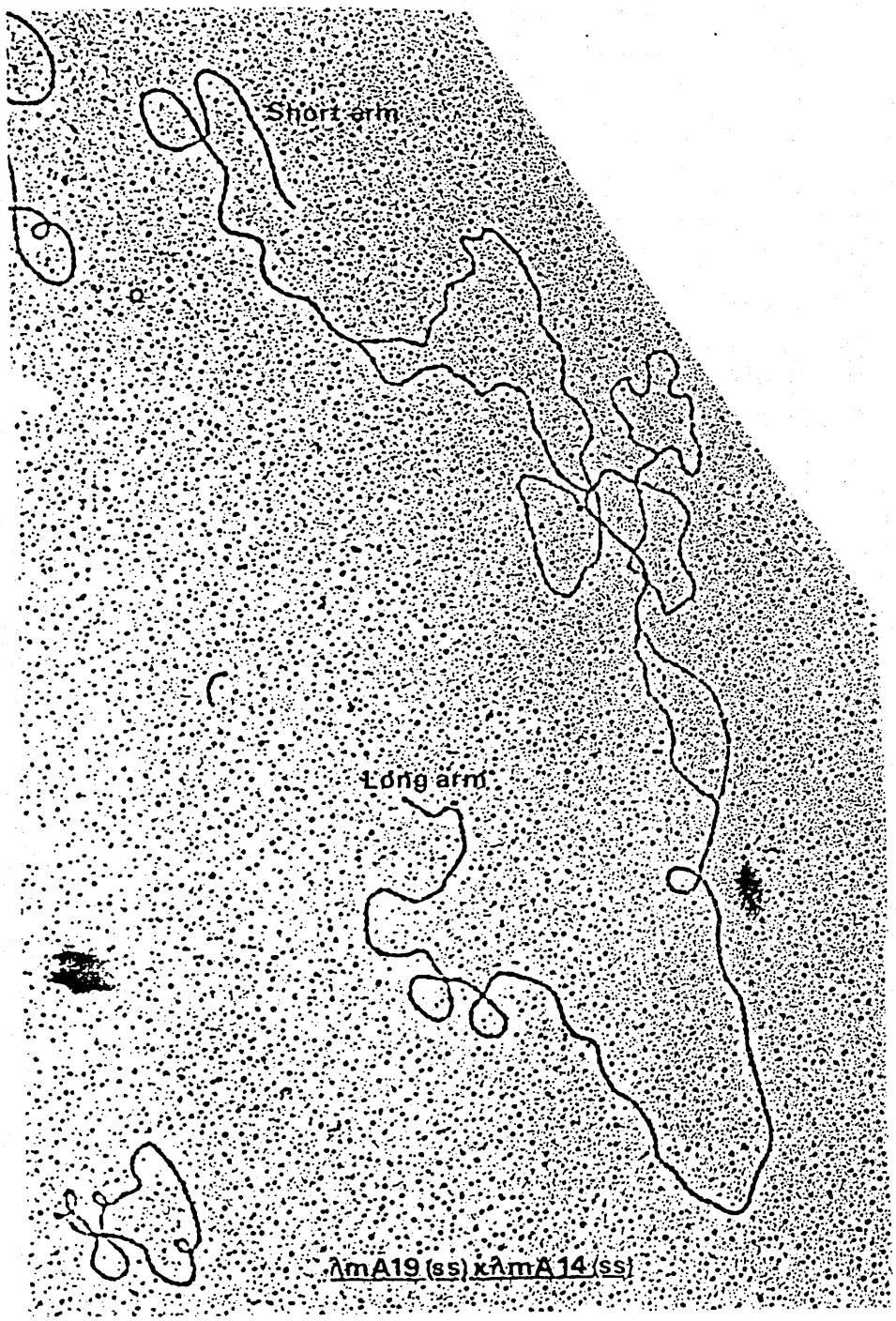
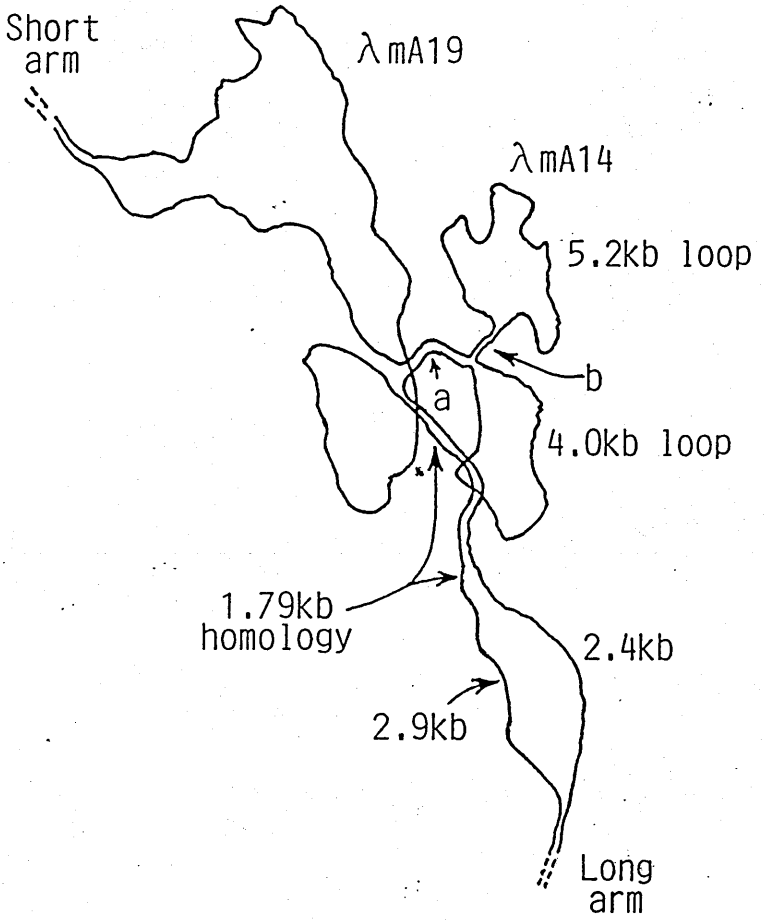


Figure 1.6 Schematic interpretation of the heteroduplex formed between separated single strands of λmA14 and λmA19

A schematic interpretation, by Dr H. Delius, of the heteroduplex shown in Figure 1.5. The electron micrograph stem sections are designated a and b.



lambda recombinants were in the same orientation relative to the arms. As the orientation of the actin-like region in λmA19 was already known to be 5' to 3' relative to the conventional representation of the long and short arms of lambda, that in λmA14 must be likewise. Measurements indicated that the heteroduplex of the actin-like region was 1.79kb in extent, and separated from the long arm of lambda by non-heteroduplex regions of 2.4 and 2.9kb. As it was already known that the actin region in λmA19 was 2.9kb from the long arm, it was concluded that the 2.4kb non-heteroduplex region represented the distance of the actin-like region of λmA14 (which must be at least 1.79kb) from the long arm of lambda. Within an estimated 50 nucleotides of the 3' end of the actin-like region, a foldback structure was observed. This foldback structure comprised a stem of 1.3kb with a 5.2kb loop at its extremity and a side loop of 4.0kb which interrupted one side of the stem. It could not be concluded from the electron micrograph whether the side loop interrupted the stem on the left or on the right-hand side. The two possibilities for the self-hybridisation structure of λmA14 are represented diagrammatically in Figure 1.7 as λmA14(a) and λmA14(b). Figure 1.8 shows the relative positions of the inverted repeat regions predicted to give rise to the structures in λmA14(a) and λmA14(b) in a linear representation with the detailed electron micrograph measurements.

The actin-like region in λmA36 was in the opposite orientation to that in λmA19 and heteroduplex analysis of λmA36 was therefore performed using λmA81. Figure 1.9 shows a schematic diagram of the electron micrograph of the heteroduplex formed. Measurements indicated that the heteroduplex between the actin-like regions was 1.74kb in extent and was separated from the short arm of lambda by non-heteroduplex regions of 2.45kb (known to be

Figure 1.7 Diagrammatic representation of the foldback structures in λmA14 and λmA36

(i) Diagrammatic representation of the two possibilities for the self-hybridising structure in λmA14, based on Figures 1.5 and 1.6.

(ii) Diagrammatic representation of the self-hybridising structure in λmA36, based on Figure 1.9.

The actin-like regions are shown as solid areas. In the case of λmA36 the actin-like region is interrupted by an estimated 540bp of extra DNA. The electron micrograph stem sections are designated a, b, c and d and can be followed by a subscript L or R, which respectively refers to the left or right-hand side of the stem.

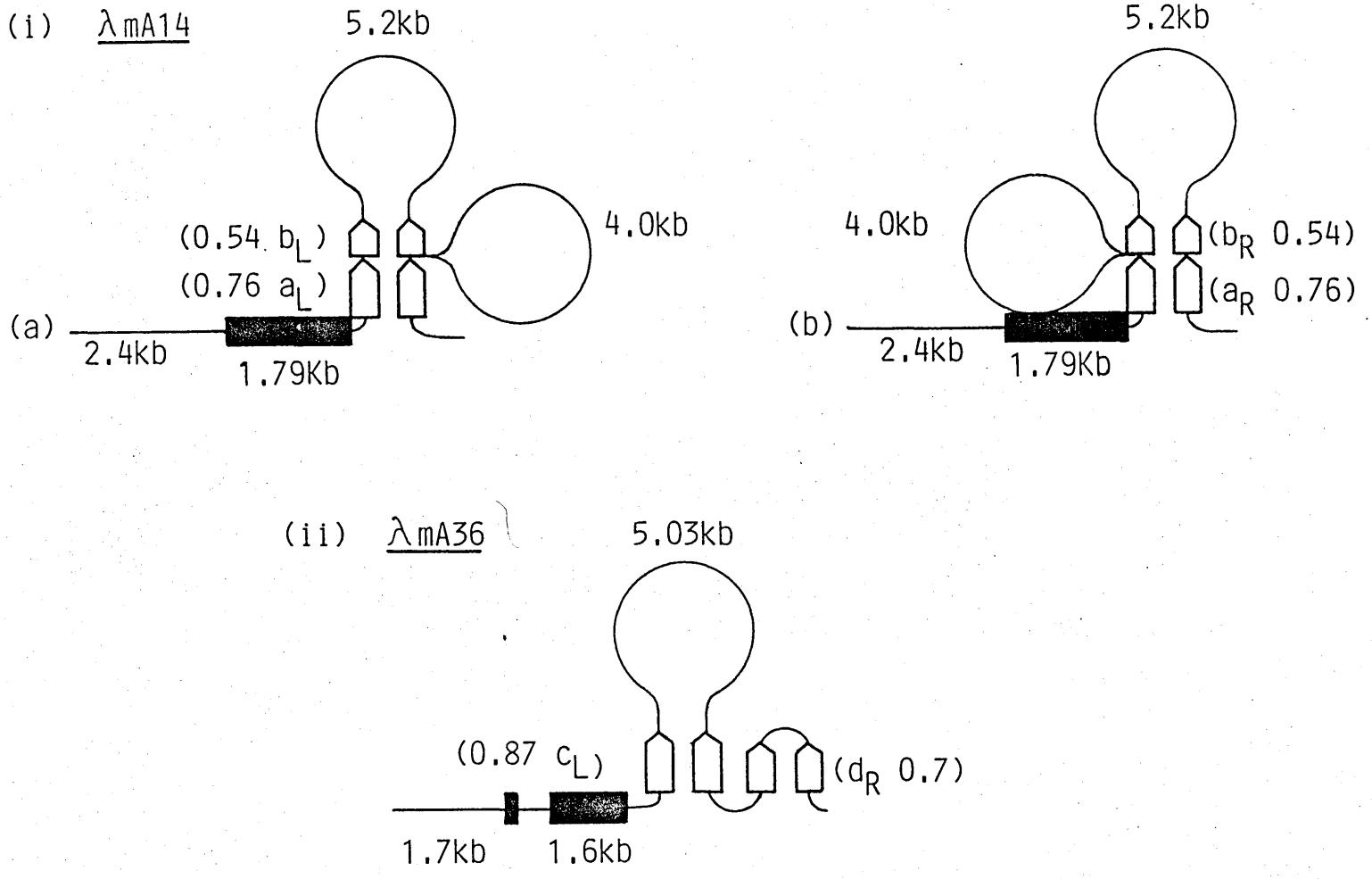

(i) λmA14
5.2kb
4.0kb
(0.54 b_L)
(0.76 a_L)
(a)
2.4kb
1.79Kb
5.2kb
4.0kb
(b_R 0.54)
(a_R 0.76)
(b)
2.4kb
1.79Kb
(ii) λmA36
5.03kb
(0.87 c_L)
(d_R 0.7)
1.7kb
1.6kb

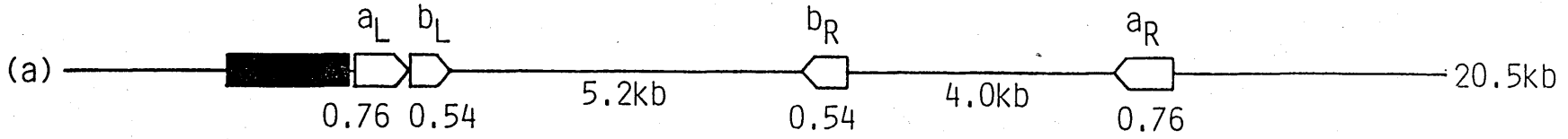
Figure 1.8 Diagrammatic representation of λmA14 and λmA36 in a linear form

(i) Shows the relative positions of the inverted repeat regions in λmA14 predicted to give rise to the foldback structures, in a linear representation of Figure 1.7, with the detailed electron micrograph measurements.

(ii) Shows the relative positions of the inverted repeat regions in λmA36 predicted to give rise to the foldback structure, in a linear representation of Figure 1.7, with the detailed electron micrograph measurements.

The actin-like regions are shown as solid areas. In the case of λmA36, the actin-like region is interrupted by an estimated 540bp of extra DNA. The electron micrograph stem sections are designated a, b, c and d which can be followed by a subscript L or R which repectively refers to the left or right-hand side of the stem.

(i) λmA14



(ii) λmA36

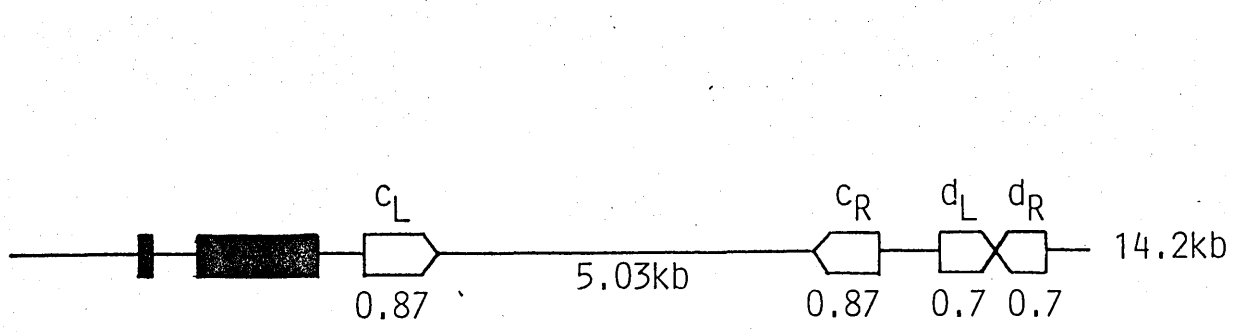
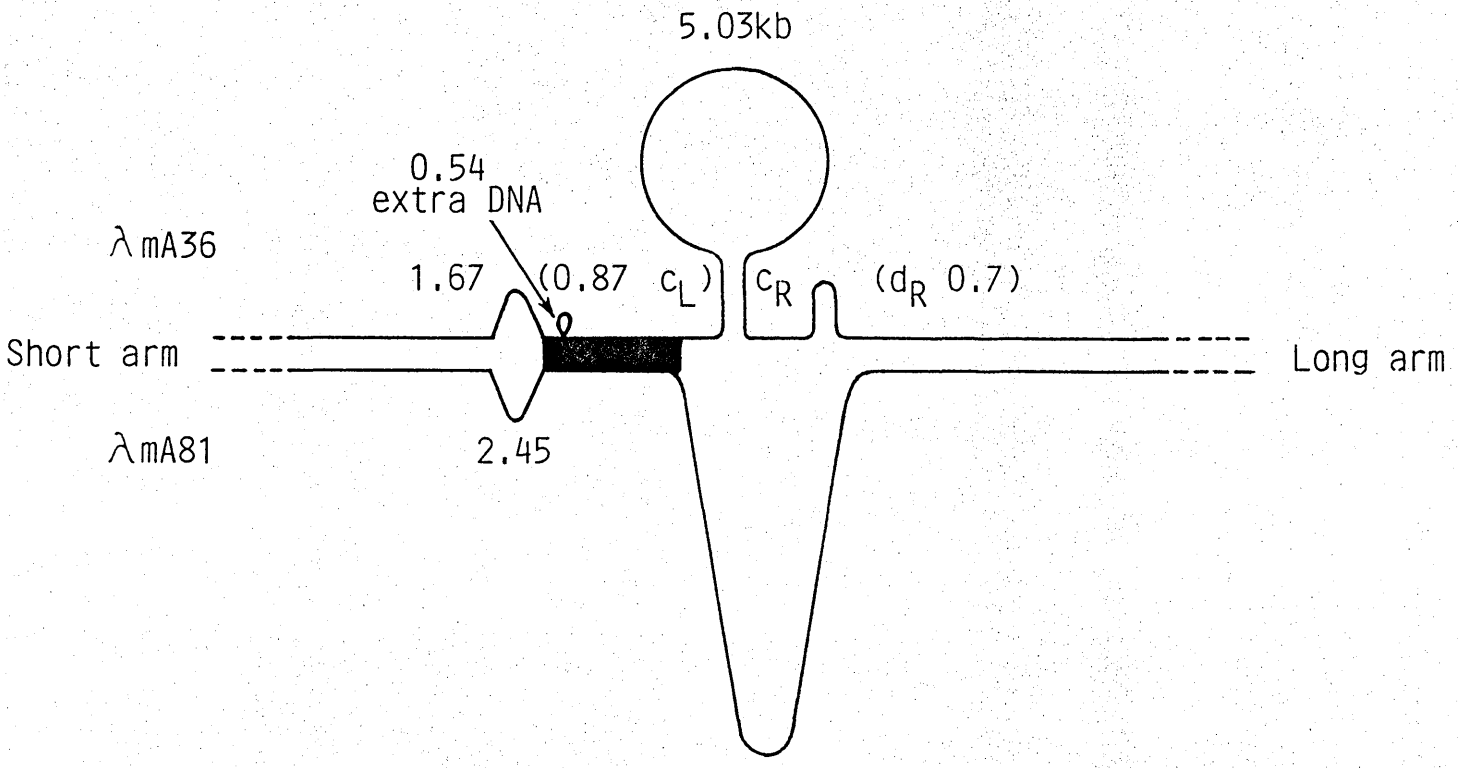


Figure 1.9 Diagrammatic representation of the heteroduplex form between separated single strands of λmA36 and λmA81

This is a schematic interpretation of the heteroduplex analysis performed by Dr H.Delius (EMBL Heidelberg). The electron micrograph stem sections are designated c and d, and can be followed by a subscript L or R which respectively refers to the left or right-hand side of the stem.


5.03kb
0.54
extra DNA
λmA36
1.67
(0.87 c_L)
c_R
(d_R 0.7)
Short arm
Long arm
λmA81
2.45

the separation in λmA81) and 1.67kb, which was concluded to be the separation in λmA36. The heteroduplex between the actin-like regions was interrupted by a 540bp region of non-homology, which was deduced to represent extra DNA, approximately 200bp from the 5' end of the actin-like region of λmA36. Within an estimated 550bp of the 3' end of the actin-like DNA in λmA36 a foldback structure was observed. The foldback structure was composed of a stem of 870bp with a loop of 5.03kb and, directly adjacent to this there was a second stem of 700bp with no loop at its end. A diagrammatic representation of the structure of the self-annealed single strand of λmA36 is shown in Figure 1.7, and Figure 1.8 shows the relative positions of the inverted repeat regions responsible for this self-hybridisation in a linear representation of λmA36 with detailed electron micrograph measurements.

Although the foldback structures of λmA14 and λmA36 both contain loops of similar size and in similar positions relative to the actin-like region, they clearly differ in detail. For example, the lengths of the heteroduplex stems were different, the foldback structure in λmA14 contained a side loop not present in λmA36, and λmA36 contained an extra stem not present in λmA14.

The precise objectives in studying the λmA14 and λmA36 were as follows. The first objective was to determine the degree of similarity between λmA14 and λmA36 over the whole of their inserts, in order to discover whether they were, in fact, related. This was of interest as processed pseudogenes are thought to arise in single independent events. The second objective was to determine the nature of the DNA which constituted the inverted repeats (foldback stems) within the two mouse genomic clones, in view of the occurrence of such structures in certain mobile elements.

CHAPTER 2 Materials and Methods

2.1 Materials

2.1.1 Chemicals

Unless otherwise specified all chemicals were Analar grade supplied by BDH Chemicals Ltd. or Fisons Scientific Apparatus. Where chemicals or equipment were obtained from other sources this is indicated in the text and a list of the names and addresses of the suppliers is given below.

2.1.2 Suppliers

Anglian Biotechnology Limited, Essex, England

Amersham International plc, Amersham, Bucks., England

Aldrich Chemical Co., Gillingham, Dorset, England

BBL Microbiology Systems, Cockeysville

BDH Chemicals Ltd., Poole, Dorset, England

Bio-rad Laboratories Ltd., Caxton Way, Watford, Herts., England

A & J Beveridge Ltd., Edinburgh, Scotland

Beckman Instrument Inc., High Wycombe, Bucks., England

Bethesda Reseach Laboratories (UK) Ltd., Cambridge, England

Bioserv Ltd., Worthing, Sussex, England

The Boehringer Corporation (London) Ltd., Lewes, E.Sussex, England

James Burrough Ltd., Fine Alcohol Division, London, England

Calbiochem-Behring Corp. (UK), Bishops Stortford, Herts., England

Cronex-Lighting, Du-pont (UK), Huntingdon, Cambs., England

Collaborative Research Inc., Universal Scientific Ltd. (UK distr), London, England

Difco Laboratories, West Molesey, Surrey, England

Fisons Scientific Apparatus, Loughborough, Leics., England

Koch-Light Laboratories Ltd. Colnbrook, Bucks., England

Kodak Ltd., Kirby, Liverpool, England

LKB Instruments Ltd., LKB House, South Croydon, Surrey, England

New England Biolabs., CP Labs. Ltd. (UK distr), Bishops Stortford, Herts., England

PL Biochemicals Inc., Northampton, England

Pharmacia Ltd., Milton Keynes, England

Schleicher and Schuell, Andermann and Co. (UK distr), East Molessy, Surrey, England

Serva, Uniscience Ltd. (UK distr), St Ann's Crescent, London

Sigma London Chemical Co. Ltd., Poole, Dorset, England

Whatman Lab Sales Ltd., Maidstone, Kent, England

Worthington, Flow Labs. Ltd., Irvine, Scotland

UV Products, Winchester, Hants., England

2.2 General Procedures

During the course of this work a number of procedures were frequently used. The following section describes these general procedures.

2.2.1 Description of bacterial strains

Three strains of bacteria have been used during the course of this project : *E.coli* Q358 (Karn *et al.*, 1980) has been used as the host for the growth of all

lambda DNA and has the following genotype :

$hsdR_k^-$, $hsdM_k^-$, supF, $\phi 80^r$, $recA^+$

Two strains of *E.coli*, JM103 and JM109 were the hosts used for the growth of all plasmid DNA. JM103 (Messing *et al*,. 1981) has the following genotype :

Δlac pro, thi, str A, supE, endA, sbcB15, hsdR4, F'traD36, proAB, $lacI^q$, ZΔM15

JM109 (Yanisch-Perron *et al.*, 1985) is a Rec A^- derivative of JM103 and has the following genotype :

rec A1, endA1, gyr A96, thi, hsd R17, sup E44, rel A1, λ^-, Δ(lac-pro AB), [F', traD36, proAB, lac $I^qZ\Delta M15$]

2.2.2 Storage of bacteria

Stocks of the bacterial strains and of the strains carrying plasmid used in this work were maintained as Hammersmith stabs (see Table 2.1). A single colony was innoculated into the stab and stored at room temperature. The bacteria remain viable for about a year under these conditions.

Frozen stock cultures of the bacteria were also maintained. 200μl of 10X Hogness freezing medium (see Table 2.1) was added to 1.8ml of an exponentially growing culture, mixed well to ensure a homogenous solution was obtained and then shock frozen in liquid nitrogen. The bacteria remain viable for several years if stored at -70^oC under these conditions.

2.2.3 Plasmid and phage

The plasmids and phage used in this study as vectors and source

Table 2.1 The compositon of the growth media

Medium	Composition per litre	
L-broth	10.0g	Bacteriotryptone (Difco 0123-01)
	5.0g	Yeast extract (Difco 0127-01)
	5.0g	NaCl
	(adjusted to pH 7.2 with NaOH)	
L-agar	1 litre	L-broth
	15.0g	Agar (Difco 0140-01)
Hammersmith agar stab	9.0g	Nutrient broth (Difco 0003-02)
	7.5g	Agar (Difco 0140-01)
	5.0g	NaCl
	10ml	10mg/ml Thymine*
10 X Hogness medium	6.3g	K_2HPO_4
	4.5g	sodium citrate
	0.9g	$MgSO_4.7H_2O$
	9.0g	$(NH_4)_2SO_4$
	18.0g	KH_2PO_4
	440.0g	glycerol
BBL-top layer agar	11.75g	Tripticase agar base (BBL 11922)
(0.65%) and $MgSO_4$	4.75g	Agar (Difco 0140-01)
	5.0g	NaCl
	10.0ml 1M	$MgSO_4$*
BBL-agar plates	11.75g	Tripticase agar base (BBL 11922)
	8.25g	Agar (Difco 0140-01)
	5.0g	NaCl

* Sterilised separately as a concentrated solution

material are listed in Table 2.2.

2.2.4 Storage of plasmid and phage DNA

Lambda and plasmid DNA was stored in TE buffer (Table 2.3) in a tight fitting capped Eppendorf tube. Plasmid DNA was stored at -20°C and lambda DNA stored at 4°C. DNA stored in this way remains stable for several years.

2.2.5 Growth media

The growth media used in the course of this work are listed in Table 2.1. All media were sterilised by autoclaving, 15lb p.s.i. for 20 min.

Any supplements to plates were added as concentrated stock solutions after the medium had cooled to 55°C and immediately before pouring.

2.2.6 Supplement to growth media

Ampicillin : The stock solution was 10mg/ml of the sodium salt of ampicillin in water. It was sterilised by passage through a 0.22µm filter (Millipore) and stored in aliquots at -20°C.

2.2.7 Commonly used solutions

During the course of this work a number of solutions were used repeatedly, Table 2.3 describes these solutions and their composition.

Table 2.2 Plasmids and bacteriophages used in this study

Plasmid	Purpose	Reference
pBR322	DNA size marker	Sutcliffe *et al.*, (1977)
pUC18	subcloning (Figure 2.1)	Yanisch-Perron *et al.*, (1985)
pmS3	cDNA probe for actin coding region (Figure 2.2)	Leader *et al.*, (1986)
pmS4-1	DNA size marker	Leader *et al.*, (1986)
MγA–ψ1	Reference clone for λmA14 and λmA36, DNA probe for actin 3'non-coding region (Figure 2.3)	Leader *et al.*, (1985)
Phage		
λ1059	Reference clone for λmA14 and λmA36	Karn *et al.*, (1980)
λcI_{857}	DNA size marker	Allet *et al.*, (1973)
λmA19	Reference clone containing a mouse γ-actin pseudogene for heteroduplex analysis	Leader *et al.*, (1985)
λmA81	Reference clone containing a mouse γ-actin pseudogene for heteroduplex analysis	Leader (unpublished)

Figure 2.1 Partial restriction map of plasmid vector pUC18

The plasmid vector pUC18 (Yanisch-Perron *et al.*, 1985), was used in the construction of the subclones in this project. This is a double-stranded circular DNA molecule, 2686bp in length. It carries a 54bp multiple cloning site (polylinker) that contains sites for 13 different restriction enzymes. The overall map shows the restriction sites of those enzymes that were used in this project. The polylinker is shown below the map. The map also shows the positions of the ampicillin resistance gene and the lac gene fragment.

Plasmid vector pUC18

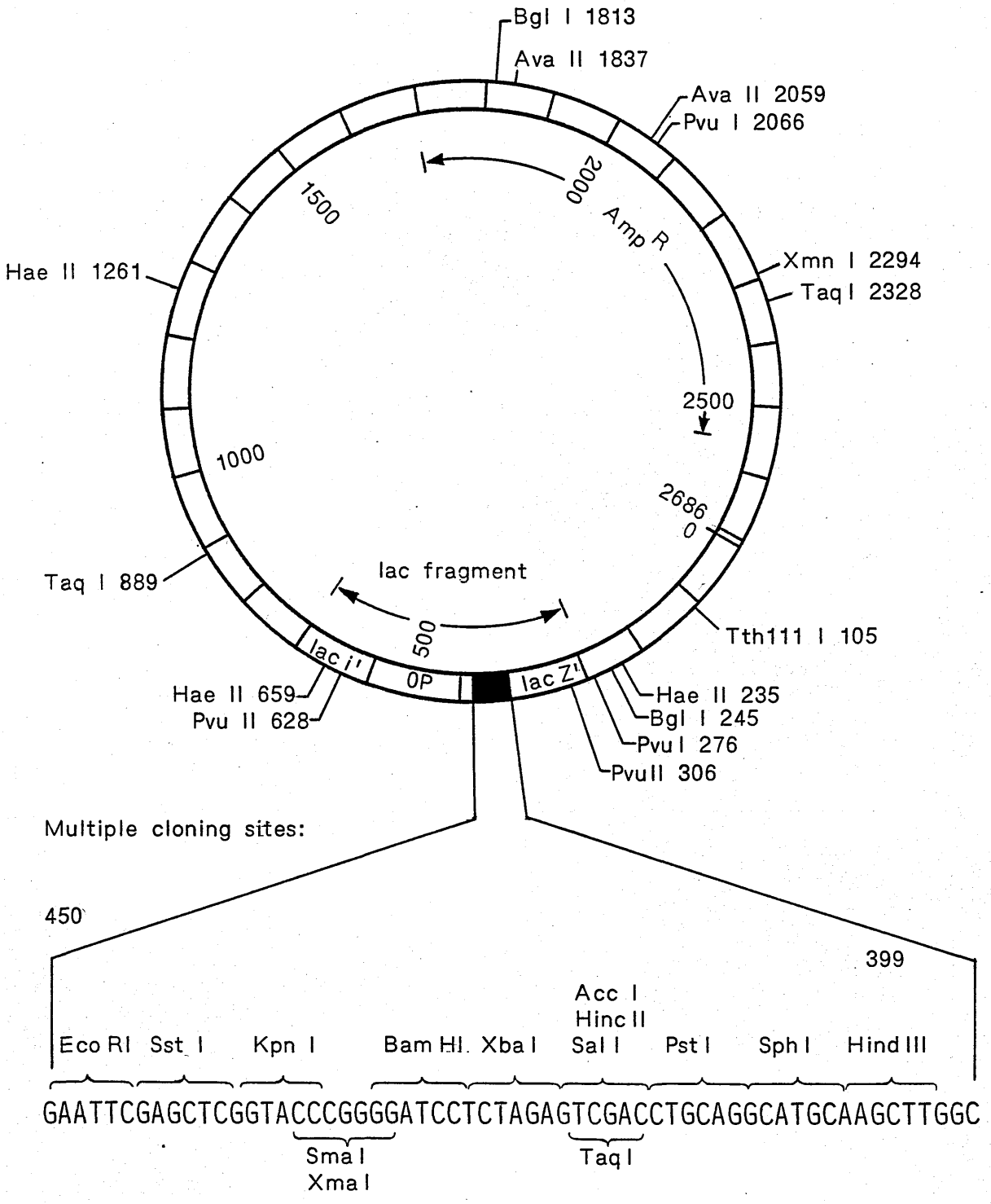


Figure 2.2 Partial restriction map of the mouse skeletal muscle actin cDNA clone pmS3

The partial restriction map of pmS3 (Leader *et al.*, 1986) is compared with the map of the corresponding mRNA. The actin coding region is represented by the solid blocks, and the 3' untranslated region is represented by the open blocks. The PstI fragment indicated, was used as an actin probe.

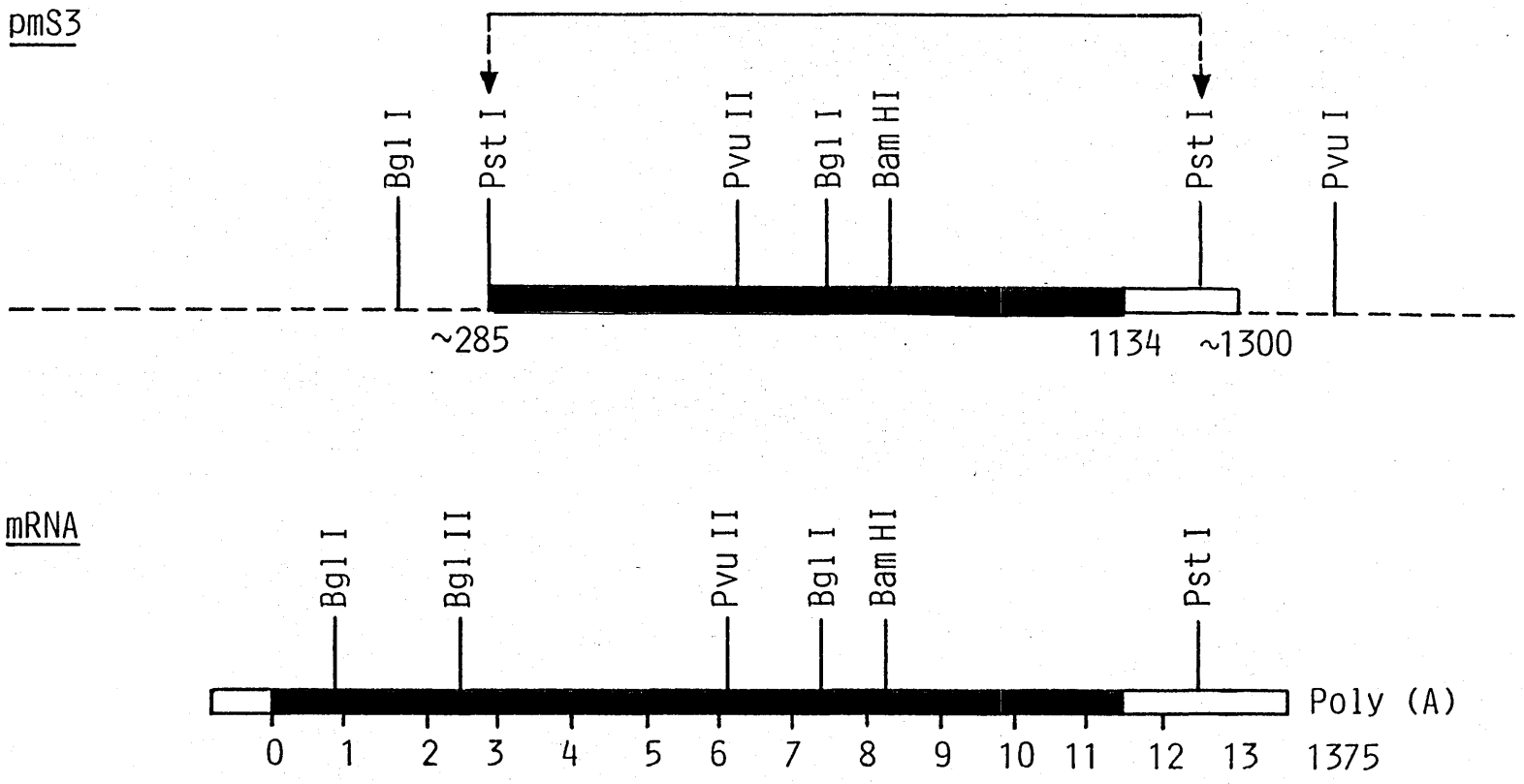

pmS3
Bgl I
Pst I
Pvu II
Bgl I
Bam HI
Pst I
Pvu I
~285
1134
~1300
mRNA
Bgl I
Bgl II
Pvu II
Bgl I
Bam HI
Pst I
Poly (A)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
1375

Figure 2.3 Partial restriction map of the actin pseudogene region within the λmA19 subclone MγA-ψ1

The plasmid subclone MγA-ψ1 contains the γ-actin processed pseudogene of λmA19 (Leader *et al.*, 1985). The partial restriction map of this subclone is only of the actin pseudogene region. The pseudo-coding region is represented by the solid blocks and the and the 3' non-coding region is represented by the open blocks.

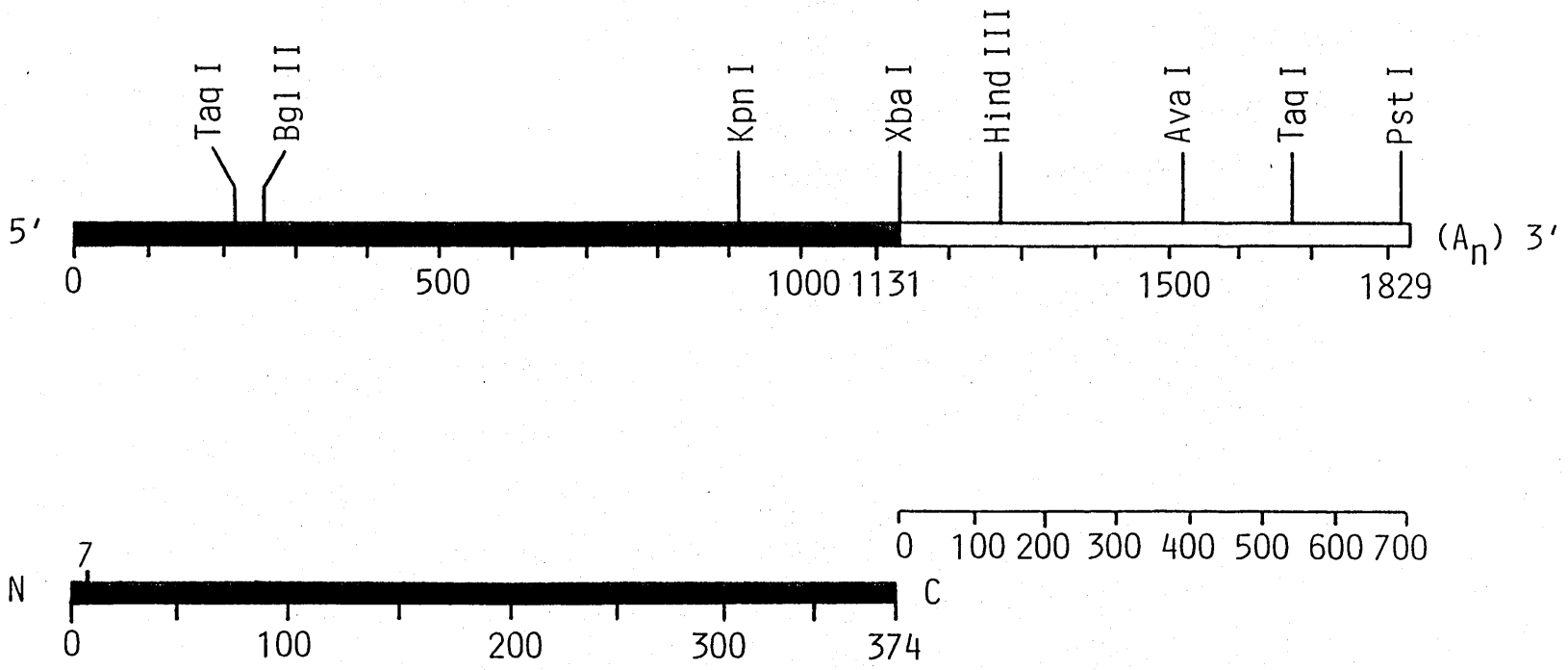

Taq I
Bgl II
Kpn I
Xba I
Hind III
Ava I
Taq I
Pst I
5′
(An) 3′
0
500
1000 1131
1500
1829
0 100 200 300 400 500 600 700
N
7
C
0
100
200
300
374

Table 2.3 Composition of commonly used solutions

Solution	Compositon	
Lambda diluent	10.0mM 1.0mM 10.0mM (* Sterilised separately as a concentrated solution)	Tris.HCl pH 7.5 EDTA $MgSO_4$*
TE	10.0mM 1.0mM	Tris.HCl pH 8.0 EDTA pH 8.0
NE	50.0mM 0.5mM	NaCl EDTA pH 7.0
10 X TBE	1.0M 0.8M 10.0mM	Tris.HCl boric acid EDTA pH 8.3
20 X SCC	3.0M 0.3M	NaCl sodium citrate
20 X SSPE	3.6M 0.2M 1.0mM	NaCl sodium phosphate EDTA
20 X SET	3.0M 0.6M 20.0mM	NaCl Tris.HCl pH 8.0 EDTA
Polyacrylamide gel elution buffer	0.5M 10.0 mM 1.0mM 0.1%	ammonium acetate magnesium acetate EDTA SDS
50 X Denhardt's solution	0.2% 0.2% 0.2% (Filter through a column of Chelex 100 ; stored at -20°C.	ficoll polyvinylpyrolidine BSA

2.2.8 Restriction digestions

Restriction enzymes were purchased from the following companies : Anglian Biotechnology Ltd., Bethesda Research Laboratories (B.R.L.), New England Biolabs and The Boehringer Corporation (London) Ltd. Enzyme digests were generally set up using one of three convenient buffers and at the temperature specified by the manufacturer.

The composition of the restriction enzyme buffers are shown below :

Buffer	NaCl	Tris	$MgSO_4$	Dithiothreitol
Low	0	10mM, pH7.4	10mM	1 mM
Med	50mM	10mM, pH7.4	10mM	1 mM
High	100mM	50mM, pH7.4	10mM	0

Restriction enzyme digests were routinely carried out in a final volume of 25µl, but larger volumes were also used where appropriate. A typical digestion mixture contained : DNA (0.5-1µg); restriction enzyme (5 units) in a final volume of 25µl restriction enzyme buffer. The mixture was incubated for 1-2 hr and the extent of digestion was monitored by electrophoresis of a small aliquot in a 1% agarose mini-gel (section 2.2.10).

2.2.9 Extraction of DNA with Phenol/chloroform and precipitation with ethanol

DNA was routinely purified free of protein by extraction with phenol/chloroform and precipitation with ethanol. Phenol was redistilled before use, saturated with TE (Table 2.3), and stored at -20°C. The extraction was carried out using phenol/TE, chloroform and isoamylalcohol in a 25:24:1 mixture which can be stored for several weeks at 4°C.

The extraction was performed as follows : the volume of the sample to be extracted was adjusted to 100µl with TE, if necessary. 100µl of the phenol mixture was added and vortexed for 3 - 4 min; then centrifuged for 1 min; the upper aqueous layer was transferred to a fresh microfuge tube and the phenol extraction procedure repeated twice more. The sample was then twice extracted with ether saturated with water to remove the residual phenol.

DNA precipitation with ethanol : The volume of the sample to be precipitated was adjusted to 100µl with TE, if necessary. A (0.1) volume of 3M sodium acetate pH 6.0 and 2.5 volumes of cold ethanol (James Burrough) was added to the DNA sample and vortexed. The samples were then placed overnight at -20°C or -70°C for 15 min; centrifuged for 10 min.

2.2.10 Agarose gel electrophoresis of DNA

DNA fragments were separated by gel electrophoresis in agarose as follows : The table below shows the concentration of agarose used, to achieve the optimum separation of DNA of various lengths.

Gel Concentration	Size Range	Recommended Voltage
0.3% Agarose	5-60kb	10V overnight
0.5% Agarose	1-30kb	40V
0.7% Agarose	0.8-15kb	40V
1.0% Agarose	0.4-8kb	60V
1.5% Agarose	0.2-4kb	60V
2.0% Agarose	0.1-2kb	80V

1% agarose was routinely used for plasmid DNA, 0.5% and 0.7% for restriction digests of lambda DNA and 0.3% agarose for genomic DNA.

Three agarose electrophoresis buffer systems were used : Loening's phosphate (Loening, 1967) and Tris.HCl borate buffers were used for routine inspection of DNA samples, and acetate buffer was used for electrophoresis of DNA where subsequent electroelution from agarose was necessary.

The phosphate electrophoresis buffer contains 36mM Tris.HCl, 30mM NaH_2PO_4, 1mM EDTA.

The acetate electrophoresis buffer contains 40mM Tris.HCl pH 7.4, 5mM sodium acetate, 1mM EDTA.

The Tris.HCl borate electrophoresis buffer contains 0.9mM Tris.HCl pH 7.4, 0.9M boric acid, 25mM EDTA.

Agarose gels were prepared by heating to boiling point the desired quantity of electrophoresis buffer containing the appropriate concentration of agarose. The agarose solution was allowed to cool to 55^{o}C, ethidium bromide (10mg/ml) was added to give a final concentration of 0.5µg/ml, and the gel

poured.

Electrophoresis was performed in a mini-gel system, gel size 12cm X 12cm, immersed in the same buffer also containing ethidium bromide (0.5µg/ml), at a constant voltage.

DNA samples were prepared for electrophoresis by the addition of 0.1 volume of dye loading buffer (1:1 glycerol : 0.025% bromophenol blue in the appropriate electrophoresis buffer).

The sizes of restriction fragments were determined by comparing with DNA marker fragments of known size, subjected to electrophoresis alongside the unknown fragments. The distances between the well and the positions where the DNA fragments of known sizes had travelled were measured and plotted on semi-log graph paper, as distance travelled (mm) against log size of DNA (kb). Similarily, the distance travelled by DNA fragments of unknown sizes were then measured, and their sizes were determined from the standard curve.

The DNA molecular weight markers routinely used were as follows : bacteriophage lambda digested with HindIII (23.7, 9.46, 6.61, 4.26, 2.26, 1.98, 0.58kb): pUC8 digested with TaqI (1443, 801 471bp) : pBR322 digested with BglI/BamHI (2319, 1288, 560, 230bp) and pMS4-1 digested with TaqI (1443, 801, 655, 383bp).

2.2.11 Polyacrylamide gel electrophoresis

Polyacrylamide gel electrophoresis was used to separate DNA fragments in preparation for sequencing by the method of Maxam and Gilbert, (1980). Vertical 160 x 160 x 1.5mm polyacrylamide gels were used and the electrophoresis buffer was 1 X TBE (Table 2.3). The concentration of acrylamide used was as follows :

Acrylamide concentration	Fragment sizes to be separated
4%	100bp and above
8%	60-400bp

The loading buffer contained 50% glycerol in the electrophoresis buffer with 0.05% xylene cyanol and 0.05% bromophenol blue as marker dyes. Electrophoresis was carried out at 200V until the dyes had travelled the required distance. In 4% acrylamide gels xylene cyanol and bromophenol blue migrates with similar mobilities to DNA fragments of 350bp and 70bp respectively. The dyes migrate at different positions in denser gels, in 8% acrylamide, xylene cyanol and bromophenol blue migrates with similar mobilities to DNA fragments of 80bp and 20bp respectively. When electrophoresis was complete the gel was removed from the apparatus and stained in a solution of ethidium bromide (0.5µg/ml) for 10 min. The DNA was visualised as described in section 2.2.12.

2.2.12 Photography of gels

DNA was visualised by ethidium bromide fluorescence on a trans-illuminator (UV Products Inc.).

Gels were photographed with a Polaroid CU-5 camera and type 665 positive/negative film.

2.2.13 Elution of DNA from acetate agarose gels

The DNA (40µg) was subjected to agarose gel electrophoresis in acetate buffer and the band of interest located using ethidium bromide staining and UV illumination. Using a scalpel, the slice of agarose containing the band of interest was cut out and placed in a dialysis bag. The gel slice was covered with acetate electrophoresis buffer and the bag tightly sealed , ensuring that no air bubbles were trapped.

The bag was immersed in a shallow layer of acetate electrophoresis buffer. After subjecting to electrophoresis for 1-2 hr, the polarity of the current was reversed for 2 min to release the DNA from the walls of the dialysis bag. The gel slice was then visualised on a UV illuminator to ensure all the DNA had been eluted from it.

All the buffer surrounding the gel slice was transferred into a 1.5ml snap-cap polypropylene Eppendorf tube and the bag was washed out with a small quantity of electrophoresis buffer. The total volume of buffer was kept down to 400µl to allow the precipitation to be performed in the Eppendorf tube, facilitating the recovery of relatively small amounts of DNA.

The buffer containing the eluted DNA was then subjected to centrifugation for 15 min to sediment any contaminating agarose debris. The supernatant was then transferred into a clean 1.5ml Eppendorf tube and precipitated with ethanol. After precipitation at -20^{o}C overnight, or at -70^{o}C for 15 min, the sample was subjected to centrifugation for 5 min, the supernatant removed and the pellet washed with 80% ethanol, chilled and recentrifuged as before. The supernatant was removed and the pellet dried under vacuum for 5 min.

2.2.14 Elution of DNA from polyacrylamide gels

The method used was based on a procedure described by Maxam and Gilbert, (1980).

After the DNA (10µg) had been subjected to polyacrylamide gel electrophoresis, the band of interest was located using ethidium bromide staining and UV illumination. Using a scalpel, the slice of acrylamide containing the band was cut out and placed in a 1ml plastic automatic pipette tip (Eppendorf type blue). The tip had been sealed at the end by heating and packed with siliconised glass wool. The polyacrylamide band was ground up using a glass rod and 600ul of elution buffer (Table 2.3), and then incubated at 37°C overnight.

The DNA of interest was eluted from the gel by rinsing the tip with 4 X 200µl elution buffer. The pooled eluate (1.4ml) was precipitated with 2.5 volumes of ethanol and left at -70°C for 30 min. The DNA was sedimented by centrifugation at 3,500 rpm for 30 min at -10°C. The DNA was suspended in 400µl 0.3M sodium acetate, transferred to an Eppendorf tube and centrifuged to remove pieces of acrylamide. The supernatant was then transferred to a new Eppendorf tube, precipitated with ethanol and dried under vacuum.

2.2.15 Blotting of DNA onto nitrocellulose

The method used was based on the procedure described by Southern, (1975).

DNA was subjected to electrophoresis through a phosphate agarose gel. In general, 0.5µg lambda phage DNA, 0.2µg plasmid DNA or 10µg genomic DNA was loaded per single gel slot. After a photographic record had been made of

the gel, the DNA was denatured by soaking the gel in 0.5M NaOH, 1.5 NaCl for 30 min. Then the DNA was neutralised by soaking the gel in 0.5M Tris.HCl pH 7.6, 1.5 NaCl for 30 min.

The DNA was then transferred to nitrocellulose (Schleicher and Schuell) using 20 x SCC (Table 2.3), overnight (16 hr) at room temperature. The nitrocellulose was then removed, washed in 2 X SCC for 5-10 min and dried on 3MM Whatman paper. Finally the filter was baked in a vacuum oven for 2 hr at 80^{o}C.

2.2.16 Preparation of a ^{32}P-labelled probes by nick-translation

When the DNA to be used was a recombinant plasmid containing inserted mammalian DNA, the insert DNA was cut out and removed from the vector to serve as a probe (2.2.13). The probe DNA was labelled by 'nick translation'. A typical labelling reaction contains the following components : Probe DNA (0.3-1µg); 50µM of each dATP, dTTP, dGTP (non-radioactive); 50µCi α^{32}P-dCTP (1mCi/100µl Amersham); DNase (10^{-7}mg/ml) and DNA polymerase (5 units) in medium restriction enzyme buffer (section 2.2.8). If a different radioactive dNTP was used, the non-radioactive dNTPs were the appropriate remaining three.

The DNase was stored as frozen stock at 1mg/ml in $H_2$0. A 1 in 10,000 dilution was made just before use.

The mixture was incubated at 15^{o}C for 4 hr. Then 100µl NE (Table 2.3) was added and the mixture applied to a Biogel P-60 (Bio-rad) column equilibriated with NE. After the sample had soaked in, the column was eluted with 9 X 100µl portions of NE, collecting each fraction separately. The peak fractions were pooled, usually fractions 5, 6 and 7, and the radioactivity

(Cherenkov radiation) of the probe determined using a scintillation spectrometer, set to the ^{3}H channel.

2.2.17 Hybridisation of ^{32}P-labelled probes onto blotted DNA

The pre-hybridisation and hybridisation reactions were performed in a polythene bag slightly larger than the nitrocellulose filter. The bag was heat-sealed with the expulsion of air.

The nitrocellulose filter was pre-hybridised in 15ml 5 X SSPE (Table 2.3), 10 X Denhardt's solution (Table 2.3), 0.1% SDS and 50% deionised formamide for 2 hr at 42°C. The formamide was deionised using mixed bed resin (Biorad).

The pre-hybridisation solution was removed and 7.5ml of fresh hybridisation solution containing denatured probe was added. The probe was denatured by adding 0.1 volume 1M NaOH for 10 min, then 0.1 volume 1M Tris.HCl pH 7.6 and 0.1 volume 1M HCl. Usually at least 10^{6} cpm of denatured probe was added per 12cm X 12cm filter. The bag was resealed and incubated overnight at 42°C.

The nitrocellulose filter was then washed in 2 X SCC (Table 2.3), 0.1% SDS for 5 X 10 min at room temperature, followed by 0.1 X SCC, 0.1% SDS for 2 X 30 min at room temperature or 45°C.

The filter was dried and exposed to Kodak-X-Omat H-film using a intensifying screen (Cronex-lighting) and left overnight at -70°C.

2.2.18 Computer programs for the analysis of DNA sequence

The following programmes were utilised in the compilation, and analysis of DNA sequences. A number of programmes devised by Staden (1978), were run on a Digital PDP 11-34 computer, with a multi-user facility in the

Biochemistry Department of the University of Glasgow. Programmes of the UWGCG (University of Wisconsin Genetics Computer Group) package (Devereux et al.,1984) were run on the EMBL (European Molecular Biology Laboratory) VAX 11/785 and VAX 8600 computers. This package contains programmes for the analysis and investigation of DNA sequences and comparison with those in the EMBL database (EMBL, Heidelberg, W.Germany).

(a) Staden programmes

SEQEDT: this program was used to create and edit a file for DNA sequences.

SEQLST: lists the sequence file created by SEQEDT in the Staden format.

TRNTRP: translates nucleotide sequences into peptide sequences in any desired reading frame using the three-letter amino-acid code.

SEARCH: searches sequences for restriction sites and strings of sequences of no more than 20 bases.

SEQFIT: searches sequence for similarities with a string of sequences less than 200 bases, and can also be used for percentage complementation.

SQRVCM: generates a sequence complementary to the sequence in question.

CUTSIT: compares given sequence file with restriction enzyme file and lists all the known restriction sites within the sequence.

(b) UWGCG programmes

FIND : searches through sequence(s) for short sequence patterns. It is able to look through large data sets for any given sequence pattern specified, recognise patterns with some symbols mismatched but not with gaps, and searches both strands of the sequence if necessary. Patterns may not be more than 41 characters long.

BESTFIT : finds the best region of similarity between two sequences, and inserts gaps to obtain the optimal alignment. The sequences can be very different lengths but the program cannot evaluate a surface of comparison larger than 10^6 base squared, with input sequences not more than 30,000 symbols long.

GAP : produces an optimal alignment between two sequences by inserting gaps in either one as necessary. It considers all possible alignments and gap positions, and creates the alignment with the largest number of matched bases and the fewest gaps.

WORDSEARCH : tries to find places where one sequence is similar to any set of other sequences. It finds segments of similarity between sequences by finding regions with an unusual number of short perfect matches, and compares both strands of the query sequence.

SEGMENTS : tries to find the best segment of similarity at the locations found by WORDSEARCH.

REPEAT : finds repeats in sequences. It allows one to choose a minimum repeat window, stringency, a search range and then finds all the repeats within these parameters.

STEMLOOP : finds stems (inverted-repeats) in nucleic acid sequences. It allows one to choose a minimum stem length, maximum loop size and minimum bonds per stem. The stems found can be sorted by position, size (stem length),

or quality (number of bonds).

(c) Other programmes

These two programmes were devised by Dr P. Taylor (Department of Virology, University of Glasgow), and were run on the Digital PDP 11-34 computer.

PHOMOL: compares two sequence files with a maximum of 2048 characters. This program uses the blocks that satify the minimum number of matches to obtain the best alignment and then align the remaining to the best. However it has limitations and sometimes misses the match.

CINTHOM: creates a homology matrix plot between two sequence files.

2.3 DNA preparations

2.3.1 Preparation of bacteriophage lambda DNA

A 50ml overnight culture of *E.coli* Q358 was sedimented using a bench-top centrifuge (Beckman) at 2,000rpm for 20 min. The supernatant was removed and the cells resuspended in 0.5 volume of sterile 10mM $MgSO_4$. A suitable amount of phage (to produce 10-100 plaques per plate) was absorbed onto 200µl of the cells in an Eppendorf tube, mixed and incubated at 37°C for 20 min. The phage and cells were then layered over 3ml BBL top-layer agar (Table 2.1) which had been cooled to 45°C, then mixed gently, and poured onto BBL plates (Table 2.1). The plates were inverted and left overnight at 37°C.

Using a sterile pasteur pipette a single plaque was removed from the BBL plate and added to 200µl of freshly saturated overnight culture of Q358 and left for 20 min at room temperature. The cells and plaque were then

transferred into a 100ml conical flask containing 20ml L-broth and 5mM $MgSO_4$. The flask was shaken at 37°C until lysis of the cells occurred. This was usually between 4 - 6 hr and is evident as the growth medium becomes clearer and bacterial debris can be seen. Chloroform (1ml) was added and the flask shaken for a further 5 min. The growth medium was then decanted into 50ml plastic tubes (Falcon), leaving the denser chloroform behind and then the bacterial debris was sedimented by centrifugation using the bench-top centrifuge 2,000rpm for 20 min. The supernatant was then titred before being transferred to a new plastic tube and stored at 4°C. The titration of the bacteriophage library procedes as described in detail above with the bacteriophage being diluted with lambda diluent (Table 2.3) to produce a range of serial dilutions between 10^{-1} and 10^{-7}. The titre of the library was calculated by counting the number of plaques per plate at any given dilution. The titre of the supernatant was about 10^{10} phage per ml.

Freshly-saturated overnight Q358 culture (7ml) was added to 1 litre of L-broth containing 5mM $MgSO_4$ in a 2 litre conical flask and shaken at 37°C until an A_{650} of 0.3 was reached. Then 5 x 10^{10} pfu was added to the 1 litre growing culture, mixed well and separated into 250ml portions each in 2 litre flasks. The flasks were shaken until lysis occurred, in approximately 3.5 hr, 2.5ml chloroform added and the supernatant decanted into large buckets and centrifuged at 4,000rpm for 20 min. At this stage the titre of phage in a sample of the supernatant was determined. Normally a value of 10^{10} phage per ml was obtained.

To the supernatant, DNase (Boehringer, grade II) and pancreatic RNase (Boehringer, grade II) were added to a final concentration of 10µg/ml. After incubating at room temperature for 30 min, solid NaCl was added to 2%, followed by the addition of PEG 6000 (Serva) to 8%. The flasks were shaken

continuously at room temperature until all the PEG 6000 had dissolved, then the flasks were left overnight at 4°C to allow the phage to precipitate.

The supernatant was centrifuged at 6,000rpm for 30 min to sediment the precipitated phage, and the pellet was then resuspended in 20ml lambda diluent (Table 2.3). After complete resuspension , 0.71g caesium chloride was added per ml to give a density of 1.5. The solution was clarified by centrifugation at 1,500rpm for 30 min and then transferred to sealable tubes (Beckman) which were centrifuged at 50,000rpm at 20°C, for 16 hr in a VTi50 rotor (Beckman).

A white band of phage particles was visible under white light and was collected by piercing the side of the tube with a hypodermic needle. The phage was further purified by centrifugation at 65,000rpm at 20°C, for 16 hr in a VTi65 rotor (Beckman).

The white phage band was collected as before and dialysed against 4 changes of 500ml 10mM Tris.HCl pH 7.5, 1mM EDTA, 10mM $MgSO_4$. The phage solution was then extracted with phenol/chloroform, precipitated with ethanol (section 2.2.9) and finally centrifuged at 10,000rpm for 10 min. The supernatant was removed and the precipitated DNA resuspended in 200-400µl TE. Boiled pancreatic RNase A (Boehringer grade I) was then added to a final concentration of 10µg/ml and left at room temperature for 30 min.

The phage DNA was stored at 4°C.

2.3.2 Preparation of bacteriophage lambda DNA from lysogenic *E.coli* M65 strain

The thermolabile strain M65 is lysogenic for bacteriophage $\lambda cI_{857}S_7$, (Allett *et.al*, 1973). The method of preparation of the lambda DNA was as

follows.

The M65 strain was first tested to ensure it was thermolable, by checking that it grew at 30°C but not at 42°C. A single colony of M65 was innoculated into 50ml of L-broth and grown overnight at 30°C.

10ml portions of the overnight culture were innoculated into four 2 litre flasks containing 200ml L-broth plus 10mM $MgSO_4$. The cultures were grown at 30°C until an A_{630} of 0.7 was reached and then transferred to a 42°C shaking water bath for 30 min. The flasks were then incubated at 37°C and shaken vigorously for 90 min. The cells were harvested by centrifugation at 6,000rpm for 15 min and then the cells were resuspended in 4ml of supernatant fluid. Chloroform (0.3ml) was added and the cell suspension shaken by hand at room temperature until the solution was very viscous. To reduce the viscosity DNase (Boehringer grade II) was added to a final concentration of 5µg/ml and incubated at 37°C for 5 min. The volume was adjusted to 20ml with lambda diluent (Table 2.3) and 14.2g caesium chloride added to give a density of 1.5. The solution was clarified by centrifugation at 1,500rpm for 30 min in a bench-top centrifuge (Beckman).

The phage was then purified by caesium chloride equilibrium centrifugation, as described in section 2.3.1.

2.3.3 Small scale isolation of plasmid DNA

The rapid method by Holmes and Quigley, (1981) was used to prepare the small scale 'mini-prep' plasmid DNA.

A single plasmid-carrying colony was streaked out onto one half of an appropriate antibiotic plate and also streaked out as a short line on a master plate. After the bacteria had grown overnight the master plate was stored

carefully away at 4°C for future reference. The bacteria on the growth plate were gently scraped off and resuspended in 1ml of lysis buffer (50mM Tris.HCl pH 8.0, 50mM EDTA pH 7.5, 8% sucrose, 5% Triton X-100) in a 1.5ml Eppendorf tube. 10µl lysozyme (20mg/ml in H_20) was added and incubated for 7 min at 95°C. The suspension was centrifuged for 15-30 min in an Eppendorf centrifuge and then 0.6ml of the supernatant was transferred to new 1.5ml Eppendorf tube. Next 2µl boiled RNase (1mg/ml) was added and incubated for 15 min at 37°C, followed by 1µl of diethylpyrocarbonate and incubation for 10 min at 65°C. Then 0.24ml 5MNH_4Ac, 0.54ml isopropanol was added, mixed well, and left on dry ice for 15 min. The DNA was precipitated by centrifugation in an Eppendorf centrifuge for 10 min, the supernatant was removed and the DNA washed with 0.3M NH_4Ac, 70% isopropanol, followed by cold ethanol. The DNA was dried under vacuum and then resuspended in 30µl of TE. The resulting concentration of DNA was generally around 1µg/ul.

Limited restriction analysis was then carried out using enzymes with known recognition sites for the recombinant of interest.

2.3.4 Large scale isolation of plasmid DNA

The method used was the alkali lysis technique of Birnboim and Doly, (1979).

The volumes given below are for a 800ml culture but were adapted for the preparation of plasmid from larger cultures or for small preparations of different plasmids.

A single colony of transformed bacteria was innoculated into 25ml L-broth containing the appropriate antibiotic and grown overnight. The

overnight culture (5ml) was then innoculated into 800ml L-broth in a 2 litre flask and incubated at 37°C with vigorous shaking until the culture reached late log phase, with an A_{650} of 0.8. Chloramphenicol (2.5ml) solution (25mg/ml in 50% ethanol) was added to final concentration of 165µg/ml and incubation was continued for a further 16 - 20 hr.

The cells were harvested by centrifugation at 5,000rpm for 5 min at 4°C. The supernatant was removed and the pellet resuspended in 4.5ml of 50mM glucose, 10mM EDTA, 25mM Tris.HCl pH 8.0 and then 0.5ml lysozyme (40mg/ml in the same solution) was added and left for 30 min at 0°C. The solution was then transferred to a 100ml polycarbonate tube, 10ml of 0.2M NaOH, 1% sodium dodecyl sulphate added, mixed well and left 5 min at 0°C. Then 7.5ml 3M NaAc pH 4.8 was added, mixed well and left 60 min at 0°C. The cell DNA and debris was pelleted by centrifugation at 30,000rpm for 30 min at 4°C.

The supernatant was divided between two 30ml corex tubes and 0.6 volume of isopropanol added to each. After allowing to stand at room temperature for 15 min, the DNA was recovered by centrifugation at 8,000rpm for 15 min at room temperature.

The DNA was resuspended in 30ml TE (Table 2.3) and transferred to a 50ml plastic tube (Falcon). The closed-circular plasmid DNA was purified from linear plasmid DNA and any remaining chromosomal DNA by centrifugation to equilibrium in a caesium chloride gradient containing ethidium bromide.

CsCl (28.9g) and 1.8ml ethidium bromide (10mg/ml) was added to the 30ml of DNA solution. The solution was transferred to a sealable tube (Beckman) and centrifuged to equilibrium at 50,000rpm for 16 - 20 hr at 20°C in a VTi50 rotor (Beckman).

Two bands of DNA were visable in the ordinary light. The upper band corresponded to the linear bacterial DNA and the nicked plasmid DNA, the

lower band consists of closed-circular plasmid DNA. The bands were more easily visualised under U.V. light (long wave) and the lower band was removed through a hyperdermic needle inserted into the side of the tube.

The ethidium bromide was removed from the ethidium bromide / DNA solution by extraction with isoamylalcohol until all the pink colour disappeared from the aqueous phase. The colourless DNA solution was transferred to a Corex tube and 4 volumes of TE added, followed by twice the total volume of ethanol. The DNA was left to precipitate overnight at -20°C.

The Corex tube was transferred to dry ice for 30 min, then centrifuged at 10,000rpm for 10 min at 0°C using the HB4 swing-out rotor of a Sorvall centrifuge. The DNA precipitate was resuspended in 100 - 200µl TE, transferred to a 1.5ml Eppendorf tube and precipitated with ethanol. After centrifugation in a Eppendorf centrifuge the DNA was washed twice with cold 80% ethanol and dried under vacuum. The DNA was dissolved in TE. The DNA concentration was determined by measuring the A_{260}, assuming that a solution of 50µg/ml DNA has an A_{260} of 1 in a cell with 1 cm light path.

From a 800ml culture a yield of 2mg of plasmid was obtained.

2.3.5 Isolation of high molecular weight DNA from mouse liver

The method used is described by Blattner *et al.*, (1978).

Six mice which had been starved overnight were killed, their livers (ca 6g total) removed quickly and dropped into liquid nitrogen. The frozen livers were ground with a mortar and pestle. Only small portions were ground at a time with the frequent re-addition of liquid nitrogen. As the liver was powdered it was added to 100ml of medium prepared as follows. To autoclaved 0.5M EDTA pH 8.0, 0.5% N-lauroyl sarcosine (Sigma); proteinase K (100µg/ml)

was added and left for 30 min at 55°C.

The mixture was incubated for 2 hr at 55°C in a rotary stirring water bath at 200rpm. The mixture was then extracted at 55°C, 3 X with phenol/ chloroform/ isoamylalcohol (25 : 24 : 1). To separate the phases, the mixture was centrifuged at 4,000rpm for 10 min at 20°C and then poured into a 250ml cylindrical separating funnel. After a few mins the phases reseparated, then the phenol phases and the interphases were run off and discarded. This was repeated for a second and third extraction, cleaning the funnel with phenol between runs.

The aqueous mixture was dialysed overnight against 4 changes of autoclaved 50mM Tris.HCl pH 8.0, 10mM EDTA, 10mM NaCl. The solution was removed from the dialysis bag into a 50ml plastic tube (Falcon) and caesium chloride added to a density of 1.7g/ml. After mixing carefully the solution was transferred to a sealable tube and centrifuged at 50,000rpm for 16-20 hr at 20°C in a VTi50 rotor (Beckman).

To collect the DNA a large bore hyperdermic needle was inserted into the side of the tube near the bottom, but above any precipitate. The fractions containing DNA were detected by their high viscosity and were collected. The DNA was dialysed overnight as before. The DNA solution was removed from the dialysis bag and precipitated with ethanol.

The yield was 3mg of high molecular weight mouse DNA.

2.4 Preparation of subclones

The following section describes the procedures involved in the construction and screening of subclones derived from recombinants of bacteriophage lambda and mouse genomic DNA.

The plasmid vector pUC18 (Yanisch-Perron, Vieira and Messing, 1985),

Figure 2.1, was used in the construction of all the subclones in this project, the DNA to be cloned being inserted into one of the unique restriction sites in the polylinker of this vector.

2.4.1 Alkaline phosphatase treatment of DNA

In order to favour the formation of hybrid molecules, the vector DNA was treated with alkaline phosphatase to remove the 5'-phosphate groups thus preventing subsequent self-ligation.

Plasmids were treated with alkaline phosphatase as follows. The plasmid DNA was cleaved with restriction enzyme(s) and purified by extraction with phenol/chloroform and precipitation with ethanol as described in section 2.2.9.

The DNA was resuspended in 20µl alkaline phosphatase buffer (50mM Tris.HCl pH 9.5, 1mM spermidine, 0.1mM EDTA) and 0.5µl calf intestinal phosphatase (70 units/µl Boehringer) added and mixed well and incubated at 37^{o}C for 30 min.

After incubation the volume of the sample was increased to 100µl using TE and then extracted with phenol/chloroform three times, extracted with ether twice and finally precipitated with ethanol. The DNA was redissolved in TE to give a concentration of 0.3µg/ul.

2.4.2 Ligation of DNA fragments

Ligation reactions were carried out in mixes containing the following : insert DNA (a suitable amount); vector DNA (0.3µg); 0.5mM ATP; 1 unit T4 DNA

ligase (Boehringer) in a final volume of 30µl ligase buffer which contains 40mM Tris.HCl pH 7.6, 10mM $MgCl_2$, 1mM dithiotheitol.

The ligation mixture was incubated overnight at 15°C. The amount of DNA in the ligation reaction was adjusted to a molar ratio of 5 : 1, insert ends : vector ends.

2.4.3 Transformation of *E.coli* by plasmid DNA and selection of recombinants

(a) Preparation of cells competent for transformation by plasmid

The bacterial stains used to make 'competent' cells were *E.coli* JM103 and JM109.

A single colony of the bacteria was innoculated into 25ml L-broth and grown overnight. An aliquot (2.5ml) of the overnight culture was transferred into 500ml of L-broth in a 2 litre flask and the cells were grown until the A_{600} reached 0.2. The cells were harvested by centrifugation at 4,000rpm for 15 min at 4°C, the supernatant removed and the cells resuspended in a total of 250ml ice-cold sterile 100mM $CaCl_2$ (half the original volume). The suspended cells were then incubated on ice for 20 min.

The cell suspension was then recentrifuged as before and the cells resuspended in a total of 5ml of ice-cold sterile 100mM $CaCl_2$. Sterile glycerol (0.5ml) was added to the cell suspension, which was then aliquoted into 1ml samples in sterile 1.5ml Eppendorf tubes. The cells were then frozen in liquid nitrogen and stored at -70°C.

These cells were viable for several months when stored at -70°C, but they could not be refrozen once thawed.

(b) Transformation of *E.coli* by plasmid DNA

An aliquot of frozen competent cells was thawed slowly on ice for 30 min. Portions of the ligation mixes (2µl and 15µl) were added to 100µl of competent cells, mixed well and left on ice for 30 min.

At least 5 min before using the antibiotic plates, 0.5% IPTG (isopropyl-thiogalactoside) in sterile water, 0.5% Xgal (5-bromo 4-chloro-3-indolyl-β-D-galactoside) in dimethyl formamide was spread over the surface of the agar plates.

After the incubation on ice the tubes were incubated at 37°C for 2 min. Then the transformation mixtures were spread over the surface of the antibiotic / IPTG / Xgal agar plates. The plates were left at room temperature until all the liquid had been absorbed, then they were inverted and incubated overnight at 37°C. Small colonies (0.1mm in diameter) appeared in 8 - 10 hr.

(c) Selection of recombinant clones on the basis of β-galactosidase activity

The puC plasmids have been constructed as cloning vectors using β-galactosidase activity as the basis of selection. The vector has a fragment of the *E.coli* lac operon containing the regulatory region and the coding information for the first 146 amino-acids of the β-galactosidase (Z) gene. The amino-terminal peptide is able to complement the product of a defective β-galactosidase gene present on the F' episome in the host cell. A 'polylinker'

DNA fragment containing several unique restriction sites for cloning has been inserted, in phase, into the amino terminal portion of the β-galactosidase gene. This insertion does not affect the complementation. However insertion of additional DNA into the 'polylinker' region generally destroys the complementation.

The complementation produces active β-galactosidase which gives rise to a blue colour when the transformed cells are grown in the presence of the inducer IPTG and of the chromogenic substrate Xgal. However when DNA is cloned into the 'polylinker' region, the β-galactosidase is inactive and the colonies appear white.

False positive white colonies occur at low frequency, probably arising through incorrect self-ligation of the vector.

(d) Identification of the desired recombinants

In order to identify bacteria which contains the recombinant plasmid of interest, several white colonies were picked and their DNA obtained. Screening of the recombinant DNA was carried out by limited restriction analysis and hybridisation to the blotted DNA with a ^{32}P-labelled probe.

The construction and identification of a recombinant plasmid of interest is described as an example below :

A 3kb mouse genomic XbaI fragment containing part of the 3' non-coding region of actin-like DNA in the genomic clone λmA36, was to be subcloned into the plasmid vector pUC18.

The DNA of mouse genomic lambda clone λmA36 (2μg) and the vector pUC18 (5μg) were digested with the restriction enzyme XbaI (section 2.2.8). The cut plasmid was then treated with alkaline phosphatase (section 2.4.1). The

mouse genomic XbaI fragments were ligated into pUC18 (2.4.2) and then transformed into JM109 'competent' cells (section 2.4.3). Thirteen white colonies were picked and their DNA obtained by the 'mini-prep' method (2.3.3).

The 'mini-prep' DNA was subjected to electrophoresis through a phosphate agarose gel, shown in Figure 2.4 The DNA was then transferred to nitrocellulose (section 2.2.15).

The 'mini-prep' DNAs (1μg) were digested with XbaI and then subjected to electrophoresis, shown in Figure 2.5. The digested DNA was also transferred to nitrocellulose.

From Figure 2.5, subclone 5X was shown to contain a 3kb XbaI fragment of λmA36. In order to confirm that 5X was the desired recombinant the two nitrocellulose filters were hybridised to a ^{32}P-labelled probe derived from the 3' non-coding region of a γ-actin pseudogene (sections 2.2.16 and 2.2.17). The autoradiographs are shown as Figures 2.4 and 2.5. The results confirm that subclone 5X contains the 3' non-coding actin region of the genomic clone λmA36.

Further limited restriction digestion anaylsis of subclone 5X indicated that it contained predicted restriction sites and the orientation of the XbaI fragment within pUC18 was determined.

2.5 Restriction mapping of recombinant lambda clones by partial digestion and hybridisation to cohesive end oligonucleotide

This is a method for the rapid restriction mapping of lambda clones developed by Rackwitz, *et al.*, (1984). Partial digestion products are selectively

Figure 2.4 Identification of the desired recombinant(s) : part I

The 'mini-prep' DNA (0.5µg), designated 1X to 13X, was subjected to electrophoresis through a 1% agarose gel (section 2.2.10). The DNA was transferred to nitrocellulose (section 2.2.15) and hybridised to a ^{32}P-labelled XbaI-PstI fragment (3' non-coding actin-like DNA) from the λmA19 subclone MγA-ψ1 (Figure 2.3).

(a) Photograph of the stained DNA gel.

(b) Autoradiograph of the nitrocellulose.

The subclones which hybridised to the ^{32}P-labelled actin probe are indicated with (*). The control plasmid was subclone 14HH1, which contains the 3' non-coding actin-like region of the genomic clone λmA14.

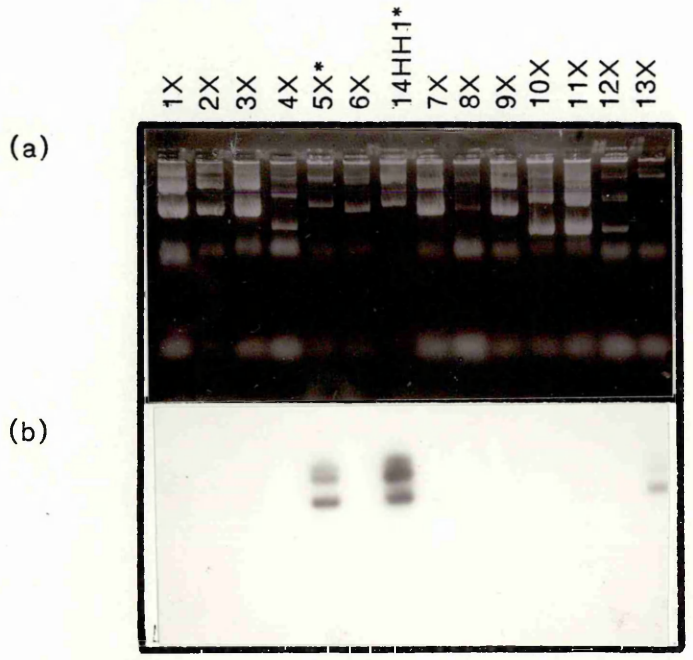

1X
2X
3X
4X
5X*
6X
14HH1*
7X
8X
9X
10X
11X
12X
13X
(a)
(b)

Figure 2.5 Identification of the desired recombinant(s) : partII

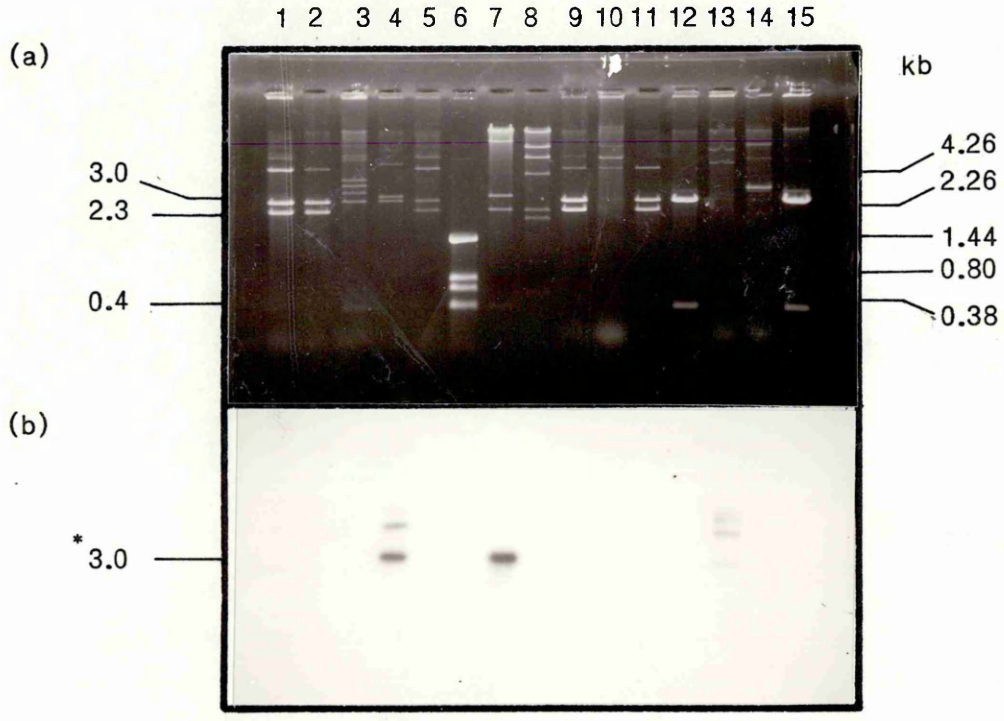
The 'mini-prep' DNA and the parent genomic clone λmA36 was digested with the restriction endonuclease XbaI and subjected to electrophoresis through a 1% agarose gel (section 2.2.10). The DNA was transferred to nitrocellulose (section 2.2.15) and hybridised to the XbaI-PstI fragment (3' non-coding actin-like DNA) from the λmA19 subclone MγA-ψ1 (Figure 2.3).

(a) Photograph of the stained DNA gel.

(b) Autoradiograph of the nitrocellulose

The length of the XbaI fragments contained the subclones and the fragments which hybridised to the ^{32}P-labelled actin probe are indicated below.

Lane	DNA	Restriction enzyme	Size of cloned XbaI fragment (kb)
1	1X	XbaI	2.3
2	2X	XbaI	2.3
3	4X	XbaI	0.4
4	5X	XbaI	3.0*
5	6X	XbaI	2.3
6	pmS4-1	TaqI	-
7	λmA36	XbaI	3.0*
8	λcI_{857}	HindIII	-
9	7X	XbaI	2.3
10	8X	XbaI	-
11	9X	XbaI	2.3
12	10X	XbaI	0.4
13	11X	XbaI	-
14	12X	XbaI	-
15	13X	XbaI	0.4


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(a)
kb
3.0
2.3
0.4
4.26
2.26
1.44
0.80
0.38
(b)
*3.0

labelled at the right or the left single-stranded cohesive end of lambda by hybridisation with the complementary ^{32}P-labelled oligonucleotide. After gel electrophoresis and autoradiography the restriction map can be read from the ladder of partial digestion products.

2.5.1 Labelling of the probe

There are two synthetic oligonucleotides complementary to the left and right cohesive ends of lambda. In this project the lambda genomic clones were mapped by selectively labelling the right cohesive end using the deoxyoligonucleotide (5' GGGCGGCGA). The oligonucleotide was labelled using the following components : 20mCi γ^{32}P-ATP (Amersham 1mCi/100µl); 10 units polynucleotide kinase (BRL); in a final volume of 10µl kinase buffer (70mM Tris.HCl pH 7.6, 10mM $MgCl_2$, 5mM dithiothreitol).

The reaction mixture was incubated at 37^{o}C for 1 hr. The percentage conversion was checked by increasing the volume of the sample to 1ml with TE and an aliquot (50µl) was separated on PEI-cellulose in 0.75M potassium phosphate pH 3.5. The oligonucleotide remains at the origin whereas ATP migrates about one third of the way to the inorganic phosphate front. The conversion was usually greater than 50%. Then the sample was heated at 100^{o}C for 1 min and stored at -20^{o}C.

2.5.2 Partial digestion and hybridisation

It was first necessary to find the digestion conditions which produced optimal partial digestion patterns required for each enzyme to be mapped. Finding the appropriate conditions was largely a matter of trial and error,

however the best method involved digesting 1µg lambda DNA with 1 unit of enzyme and stopping the reaction at suitable time points between 2 - 60 min with the addition of 20mM EDTA. Different time points were mixed to achieve a full representation of partial digestion products.

The radioactive probe was then hybridised to the partial digestion products. ^{32}P-labelled probe (2µl), which represents about 200,000cpm was added to the DNA sample which was mixed well and incubated at 75°C for 2 min followed by 2 hr at 37°C.

2.5.3 Gel electrophoresis and autoradiography

The best results were achieved using a large electrophoresis apparatus, in which the samples were separated out in 0.5% agarose for 24 hr at 1.5V per cm. The gel was then dried onto Whatman DE-81 cellulose paper followed by autoradiography as described in section 2.2.17.

The position at which different restriction enzyme sites occur along the lambda recombinant can be read directly from the autoradiograph.

2.6 DNA sequencing by the Maxam and Gilbert chemical method

Cloned DNA was sequenced by the method of Maxam and Gilbert, (1980).

2.6.1 5' end and blunt end labelling

After the DNA to be sequenced has been digested with an enzyme to generate a 5' protruding end, the end can either be labelled by the filling in reaction of the Klenow fragment of DNA polymerase with the appropriate

labelled α^{32}P-dNTP or by replacing the 5' phosphate group of the DNA using polynucleotide kinase and γ^{32}P-ATP.

Blunt ends can also be labelled by replacing the 5' phosphate groups of the DNA, however the efficiency with which the polynucleotide kinase achieves this is much lower than that for the 5' protruding ends.

(a) The Klenow reaction

To the lyophilised DNA fragment (5µg) the following components were added : 50µCi α^{32}P-dATP (Amersham 1mCi/100µl) or the appropriate radioactive dNTP ; 4µM of each non-radioactive dTTP, dCTP, dGTP ; 2 units Klenow fragment (Boehringer) and (3.75µl) 10 X medium restriction enzyme buffer (2.2.8) in a final volume of 25µl.

The reaction mixture was incubated at room temperature for 30 min and then 90µl 2.5M NH_4Ac, 360µl cold ethanol was added, mixed well and precipitated in dry ice for 5 min. The sample was centrifuged for 5 min in an Eppendorf centrifuge, the supernantant removed, and 100µl 0.3M NaAc pH 6.0, 300ul cold ethanol added. The DNA was precipitated as before and the pellet washed with cold 80% ethanol before drying under vacuum.

(b) Phosphatase reaction

To achieve 5' end labelling of a DNA fragment with γ^{32}P-ATP the 5' phosphate groups must be removed. The method to remove these phosphate groups is outlined below.

The DNA fragment was dissolved in 100µl of TE and 0.5µl calf intestinal alkaline phosphatase (70units/ul BRL) was added. After incubation at 37°C for 60 - 75 min the DNA sample was extracted with phenol saturated with TE. The phenol phase was re-extracted with an equal volume of TE. The aqueous phases were pooled and residual phenol removed by extraction with ether saturated with water. The DNA sample was finally precipitated with ethanol.

(c) The polynucleotide kinase reaction

If the DNA fragment (5µg) to be labelled had 5' protruding ends the reaction mixture contained the following components : 5µM dithiothreitol ; 60µCi γ^{32}P-ATP (Amersham 1mCi/100µl); 5 units polynucleotide kinase (PL Biochemicals) in a final volume of 11µl kinase buffer (50mM Tris.HCl, pH 8.0, 10mM $MgCl_2$).

The components were mixed well and incubated at 37°C for 30 min, then 40µl 2.5M NH_4Ac, 160µl cold ethanol was added. The DNA was precipitated on dry ice for 15 min and centrifuged in an Eppendorf centrifuge. The method was completed as described in this section part (a).

If the DNA fragment (5µg) to be labelled had blunt ends, the following conditions were used :

To the dried DNA the following components were added: 1mM spermidine ; 60µCi γ^{32}P-ATP (Amersham 1mCi/100µl); in kinase buffer (50mM Tris.HCl pH 9.5, 10mM $MgCl_2$).

The mixture was heated at 90°C for 2 min and then chilled on ice. Then two more components were added : 1mM dithiothreitol ; 5 units polynucleotide

kinase (PL Chemicals).

The method was continued as described for the 5' protruding ends.

(d) Separation of labelled fragments

The 5' labelled ends of a piece of double-stranded DNA are separated by cleavage of the fragment into two or more subfragments using a restriction enzyme which is known to cut within the DNA fragment, followed by polyacrylamide gel electrophoresis (section 2.2.11). The desired bands once visualised by ethidium bromide staining of the polyacrylamide gel, are cut out and the DNA eluted from the polyacrylamide by the method described in section 2.2.14.

2.6.2 Base-specific chemical cleavage reactions

(a) Solutions

1. Pyridine Formate : 4% v/v adjusted to pH 2.0 with pyridine (using 0.005M H_2S0_4 as a pH 2.0 standard). Stored at 4^oC.

2. DMS Buffer : 50mM sodium cacodylate, 10mM $MgCl_2$, 0.1mM EDTA pH 8.0. Store at 4^oC.

3. 'DMS Stop': 1.5M sodium acetate, 1M 2-mercaptoethanol (Koch-Light), 100µg/µl yeast RNA. Stored at -20^oC.

4. 'Hydrazine stop' : 0.3M sodium acetate, 0.1mM EDTA, 50µg/ml yeast RNA. Stored at 4^oC.

(b) Additional reagents

1. Dimethylsulphate - DMS (Gold Label, Aldrich Chemical Co. Ltd.)
2. Hydrazine - HZ (Kodak Ltd.)
3. Piperidine (Koch-Light)

(c) Base modification reactions and chain cleavage

The four reactions used for full sequence determination were specific for guanine (G), guanine and adenine (G+A), cytosine (C) and cytosine and thymine (C+T). Chain cleavage was achieved using 1M piperidine. The precise procedure followed for each of the four reactions was as follows.

Calf thymus carrier DNA (4µg) and 11µl H_2O was added to the lyophilised ^{32}P-labelled DNA fragment (1µg). The DNA sample was mixed well and then divided equally into four Eppendorf tubes labelled G, A(+G), T(+C) and C. Each tube then received different components : 98µl DMS buffer into tube G: 11µl H_2O into tube A(+G): 6µl H_2O into tube T(+C) and 8µl H_2O saturated with NaCl into tube C.

Pyridine formate (2.5µl) was added to tube A(+G), mixed and then incubated at 30°C for 70 min. The reaction was stopped by freezing the sample at -70°C for 5 min followed by drying under vacuum. The sample was washed with H_2O and dried as before.

DMS (0.5µl) was added to tube G, mixed and incubated at 20°C for 5 min. The reaction was stopped by the addition of DMS-stop (24µl) and cold ethanol (400µl) and then left for 15 min at -70°C.

HZ (15µl) was added to tubes T(+C) and C, mixed and incubated at 20°C. The reactions were stopped in tubes T(+C) and C after 8 and 10 min respectively, with the addition of HZ-stop (60µl) and cold ethanol (250µl) and then left at -70°C for 15 min.

Tubes G, T(+C) and C were centrifuged for 5 min in an Eppendorf centrifuge, the supernatant removed and the DNA precipitated with 0.3M NaAc and ethanol. The DNA was recentrifuged, the supernatant removed , the DNA washed with 70% ethanol and then dried under vacuum.

1M Piperdine (100µl) was added to all four tubes G, A(+G), T(+C) and C. The samples were mixed well, heated at 90°C for 30 min and then frozen at -70°C and dried under vacuum overnight (16 hr). To remove the residual piperidine the samples were washed twice with water and dried under vacuum.

2.6.3 Gel electrophoresis

High resolution thin (0.4mm) sequencing gels were used according to Sanger and Coulson, (1980). 6% polyacrylamide gels were routinely used which contain 7M urea and electrophoresis was in 1 X TBE buffer (Table 2.3).

The gel was subject to pre-electrophoresis at 25-30mA for 1-2hrs (LKB 2103 power pack). During this time the samples were dissolved in sequencing loading dye (99% deionised formamide, 0.05% xylene cyanol). 10,000cpm (Cherenkov) per loading was sufficient for an overnight exposure, so when possible the DNA sample was dissolved in an appropriate volume of loading dye to give 10,000 cpm per µl.

When the gels were ready, the DNA samples were boiled for 2 min then quickly chilled on ice. Three consecutive loadings were carried out per gel

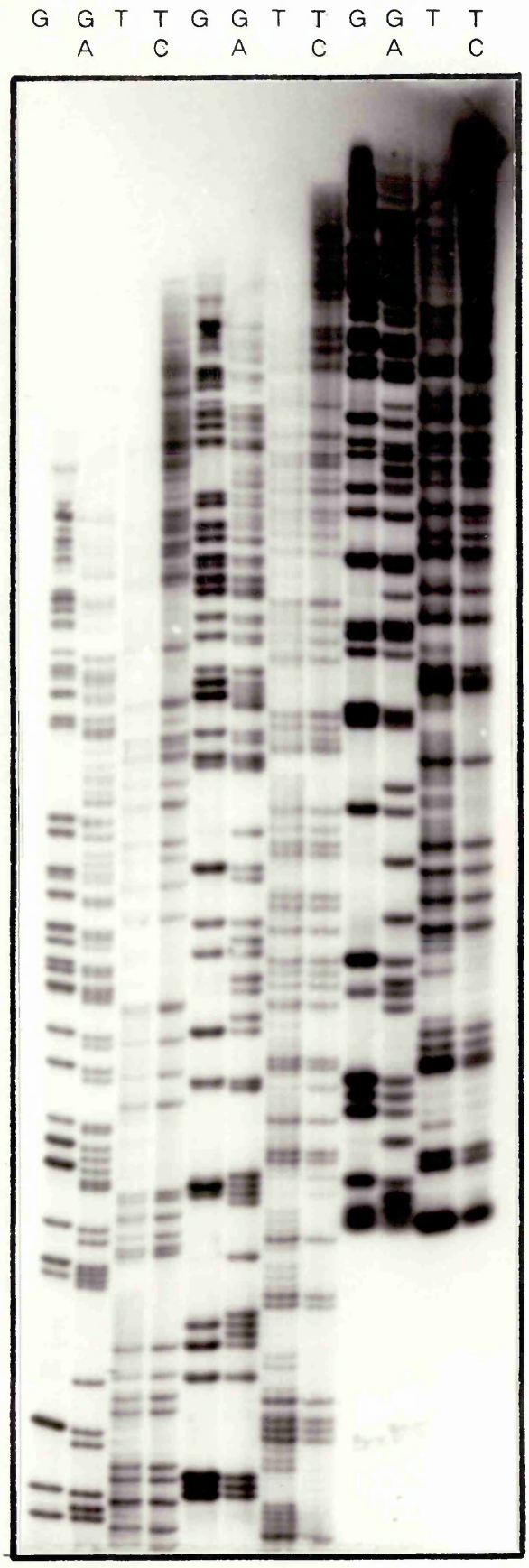
and usually a 6% polyacylamide gel allowed up to 200 nucleotides to be read from the labelled end.

2.6.4 Autoradiography

One of the glass plates was removed to expose the gel which was carefully covered with cling film. The gel was exposed to a sheet of Kodak-X-Omat H-film with an intensifying screen (Cronex-Lighting) at -70^{o}C. The gel was exposed for 1-7 days, depending on the amount of radioactivity loaded . Figure 2.6 shows an example of an autoradiograph of a sequencing gel.

Figure 2.6 An example of a DNA sequencing gel by the method of Maxam and Gilbert

The λmA14 HindIII-SstI subclone, 14HH4A (Figure 3.11) was restricted with EcoRI, 5' Klenow end labelled and secondary cleaved with HindIII. Maxam and Gilbert sequencing was performed from the EcoRI site, (a polylinker restriction site of pUC18) and then the radioactively labelled DNA fragments were separated by polyacrylamide gel electrophoresis, allowing determination of the nucleotide sequence of gel run number 3, in Figure 3.36.


G G T T G G T T G G T T
A C A C A C

CHAPTER 3 RESULTS

3.1 Determination of the similarity betweem λmA14 and λmA36

As already described in the Introduction, the two genomic clones, λmA14 and λmA36, containing the actin-like genes, were each shown by electron microscopy to be associated with DNA capable of forming large foldback structures, which although distinct in appearance shared certain similarities. The first section of this chapter describes experiments to determine the extent of similarity between these two clones.

3.1.1 Restriction endonuclease mapping of λmA14 and λmA36

The first and major experimental approach adopted to compare the two genomic clones was to construct restriction maps of λmA14 and λmA36. The sizes of the mouse DNA inserts in these recombinants were taken from the electron micrograph heteroduplex measurements, 20.5kb in the case of λmA14 and 14.2kb in the case of λmA36. Although the actin-like regions in λmA14 and λmA36 are in the opposite orientation with respect to the conventially designated left-hand (long) and right-hand (short) arm of lambda, it was clearly necessary to represent them in the same orientation for comparison, the orientation of λmA36 being the one which was reversed. This is a potential source of confusion if the terms right and left-hand are used in discussing these maps, therefore reference will instead be made to the arms of the lambda vector as 'long' and 'short'. Although positions within

the inserts of the genomic clones are frequently indicated to be 5' and 3' with respect to the actin-like sequence, it has not always been possible to avoid the use of right and left-hand. Where this occurs, it is always in relation to the common representation presented in the figures.

The genomic clones λmA14 and λmA36 were digested with several restriction enzymes as illustrated in Figure 3.1, and the fragments produced by single restriction digestion are listed in Table 3.1. The restriction enzymes used were selected on the basis that they cleave mammalian DNA relatively infrequently and also, in most cases, have no or very few recognition sites along the lambda arms of the parent vector λ1059. The ease by which restriction sites were mapped within λmA14 and λmA36 depended on the complexity of the digestion pattern produced. For example, the restriction enzyme SstI produces a relatively simple digestion pattern for both genomic clones. With λmA14 it produces four fragments of lengths 24.3, 14.5, 7.5 and 3.2kb. As the vector has no SstI sites, the three SstI sites must occur within the mouse DNA of this recombinant. The long and short arms of the vector λ1059 are respectively, 20 and 9kb, in length and therefore these arms must be contained within the 24.3 and 14.5kb SstI fragments, respectively. The order of the two internal SstI fragments (of lengths 7.5 and 3.2kb) could not be determined from this information alone. Single digestion with enzymes such as AvaI, BglII and PvuII, which cleaved several times in the vector arms and in the mouse DNA, was less informative at this stage. However, as will emerge below, the data accumulated from digestion with these enzymes was useful when combined with those from other methods.

Even from these initial results, it was evident that λmA14 and λmA36 shared some restriction fragments of similar size, for example, 2.5 and 0.4kb BamHI fragments, and 4.3 and 4.7kb BglII fragments. This reinforced the

Figure 3.1 Single restriction enzyme digestion of λmA14 and λmA36

The mouse genomic clones λmA14 and λmA36 and the parent vector λ1059 were digested with the restriction endonucleases indicated (section 2.2.8). The DNA was subjected to electrophoresis through a 0.7% agarose gel (section 2.2.10), and the molecular weight marker is λcI_{857} digested with HindIII.

	Lane	DNA	Restriction Enzyme		Lane	DNA	Restriction Enzyme
(a)	1	λmA36	AvaI	(b)	1	λmA36	KpnI
	2	λmA14	AvaI		2	λmA14	KpnI
	3	λ1059	AvaI		3	λ1059	KpnI
	4	λmA36	PvuII		4	λmA36	SstI
	5	λ1059	PvuII		5	λmA14	SstI
	6	λmA14	PvuII		6	λ1059	SstI
	7	λcI_{857}	HindIII		7	λcI_{857}	HindIII
	8	λmA36	HindIII		8	λmA36	BglII
	9	λmA14	HindIII		9	λmA14	BglII
	10	λ1059	HindIII		10	λ1059	BglII
(c)	1	λmA36	EcoRI	(d)	1	λmA36	BamHI
	2	λmA14	EcoRI		2	λmA14	BamHI
	3	λ1059	EcoRI		3	λ1059	BamHI
	4	λmA36	XbaI		4	λmA36	PstI
	5	λmA14	XbaI		5	λmA14	PstI
	6	λ1059	XbaI		6	λ1059	PstI
	7	λcI_{857}	HindIII		7	λcI_{857}	HindIII
	8	λmA36	SalI		8	λmA36	SmaI
	9	λmA14	SalI		9	λmA14	SmaI
	10	λ1059	SalI		10	λ1059	SmaI

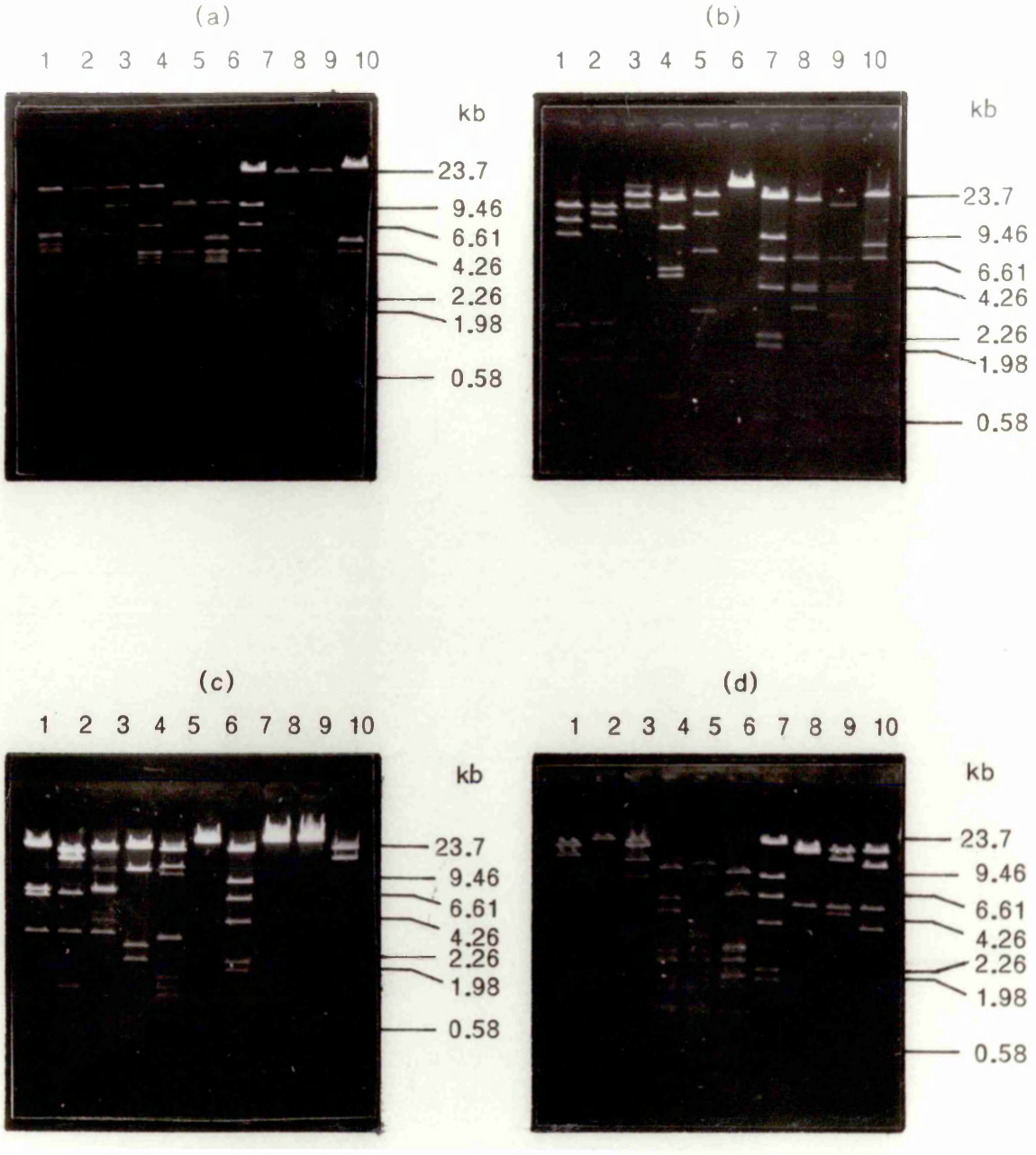

(a)
1 2 3 4 5 6 7 8 9 10
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
(b)
1 2 3 4 5 6 7 8 9 10
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
(c)
1 2 3 4 5 6 7 8 9 10
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
(d)
1 2 3 4 5 6 7 8 9 10
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58

Table 3.1 Fragments produced by single restriction digestion of λmA14 and λmA36

The lengths of the fragments produced by single restriction digestion of λmA14 and λmA36 with each endonuclease was determined as described in section 2.2.10. Similar sized fragments corresponding to the insert DNA, but not that of the vector, are boxed. The fragments which are labelled with a (*), hybridised to the ^{32}P-labelled PstI fragment of the skeletal muscle cDNA clone, pmS3 (Figure 2.2).

Restriction Enzyme	λmA14	λmA36	Restriction Enzyme	λmA14	λmA36
AvaI	14.7	14.7	KpnI	17.0	17.0
	5.9	5.6		15.3*	12.4*
	5.6	5.6		11.0	9.8
	4.9	5.3*		2.7	2.5*
	4.7	4.7		2.0*	1.5
	3.8*	4.6		1.5	
	3.5	1.9			
	1.9	0.8	PstI	13.5	11.5
	1.8			10.0*	6.0
	Many fragments < 1.0kb			3.5	4.8*
				Many fragments < 3.0kb	Many fragments < 3.0kb
BamHI	26.7*	20.4			
	9.0	15.7*	PvuII	10.0	13.0*
	4.4	2.6		5.0	6.0
	2.6	2.6		5.0	4.3
	2.6	1.5		4.3	4.2
	1.9	0.4		4.2	3.9
	1.9			3.9	3.6
	0.4			3.7*	1.7
				3.6	1.6
BglII	22.0	22.6		1.7	
	7.0	7.0		1.5*	
	4.7*	4.7*		Many fragments < 1.0kb	Many fragments < 1.0kb
	4.3	4.3			
	3.0	3.5	SmaI	19.5	19.5
	2.2			16.2*	17.7*
	2.0			6.0	6.0
	Many fragments < 1.0kb	Many fragments < 1.0kb		5.5	
				1.8	No sites
EcoRI	21.8	24.8		0.5	
	15.0	7.5*			
	7.0*	7.0	XbaI	23.5*	25.0
	3.5	3.5		13.4	12.5*
	1.6	0.4		3.5	3.1
	0.4			1.8	2.2
				1.6	0.4
HindIII	23.6*	22.0		1.4	
	6.6	8.5*		0.5	
	6.5	5.3		0.5	
	4.1	4.1			
	3.0	2.3	SstI	24.3*	20.5
	2.3	1.0		14.5	11.3
	2.3			7.5	5.5
	1.0			3.2	5.2*
					0.8

impression suggested by the initial electron micrographs that there might be similarities between these clones extending outwith the actin region.

Figure 3.2 shows the very limited partial restriction maps of λmA14 and λmA36 using only the results from the single restriction digestions. The XbaI and HindIII sites in both λmA14 and λmA36 fall within the region where the electron microscopic heteroduplex measurements predicted the actin DNA to be located. This was consistent with the occurrence of XbaI and HindIII sites in the actin processed pseudogene region of λmA19 (Leader *et al.*, 1985), the location of which has been indicated in Figure 2.3 for reference .

The results of further experiments in which the products of single restriction digestion were hybridised against ^{32}P-labelled actin probes, when considered in the context of the electron microscopic assignment of the position of the actin pseudogene, provided further information and confirmed some of the conclusions already reached. Figure 3.3 is an example of a single restriction enzyme digestion of λmA14 and λmA36, hybridised against a ^{32}P-labelled PstI fragment of the skeletal muscle cDNA clone, pmS3 (Figure 2.2), which contains DNA predominately from the coding region. Because of the high conservation of amino-acid sequence in different actin isoforms (see Introduction) this probe will hybridise to the restriction fragments containing actin DNA, even if it is related to a different isoform. The hybridising restriction fragments are indicated in Table 3.1. The actin probe hybridised to single SstI, XbaI and HindIII fragments of λmA14, which were approximately 23.0kb in length, confirming that the actin-like coding DNA within λmA14 occurred to the left of these sites. This was still consistent with the position of the XbaI and HindIII sites in λmA19, (Figure 2.3), as these sites in λmA19 are respectively, at the start of, and within the 3' non-coding

Figure 3.2 Partial restriction maps of λmA14 and λmA36 (version I)

Very limited partial restriction maps of λmA14 and λmA36 were constructed using only the results from the single restriction digestion (Table 3.1).

The position of the actin pseudogene regions, predicted from electron microscopy, are shown alongside the maps, solid areas being the pseudo-coding region and open areas indicating the 3' non-coding region. In the case of λmA36, the pseudogene coding region is interrupted by an estimated 540bp of extra DNA.

λmA36
Short arm
Eco RI
Sst I
Xba I
Hind III
Bam HI
Xba I
Eco RI
Hind III
Sst I
Bam HI
Long arm
1.0kb
λmA14
Long arm
Eco RI
Xba I
Hind III
Sst I
Bam HI
Sst I
Xba I
Hind III
Bam HI
Short arm

Figure 3.3 Example of products of single restriction digestion of λmA14 and λmA36 hybridised to ^{32}P-labelled actin probe

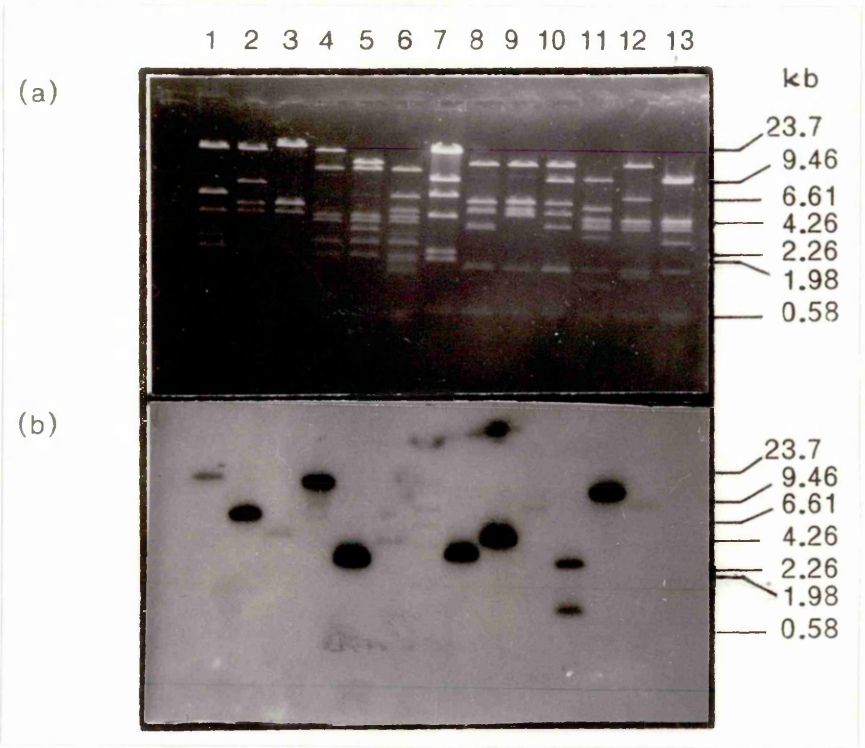
λmA14 and λmA36 were digested with the restriction endonucleases indicated (Figure 2.2.8) and subjected to electrophoresis through a 0.7% agarose gel (section 2.2.10). The DNA was transferred to nitrocellulose (section 2.2.15) and hybridised to a ^{32}P-labelled PstI fragment of the skeletal muscle cDNA clone, pmS3 (Figure 2.2).

(a) Photograph of the stained gel

(b) Autoradiograph of the nitrocellulose

The fragment(s) which hybridised to the actin probe are indicated below :

Lane	DNA	Restriction Enzyme	Hybridised fragment(s) (kb)
1	λmA14	HindIII	23.6
2	λmA36	HindIII	8.5
3	λ1059	HindIII	-
4	λmA14	ClaI	23.0
5	λmA36	ClaI	3.0
6	λ1059	ClaI	-
7	λcI_{857}	HindIII	-
8	λmA14	AvaI	3.8
9	λmA36	AvaI	5.3
10	λ1059	AvaI	-
11	λmA14	PvuII	4.3 and 1.5
12	λmA36	PvuII	13.0
13	λ1059	PvuII	-


1 2 3 4 5 6 7 8 9 10 11 12 13
(a)
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
(b)
23.7
9.46
6.61
4.26
2.26
1.98
0.58

actin region, and thus would be expected to give rise to only a single fragment, hybridising to a probe from the actin coding region. Single 12.5kb XbaI and 6.5kb HindIII fragments of λmA36 hybridised to the actin probe, confirming the position of these sites relative to the actin-like DNA, as shown in Figure 3.2. The BglII and KpnI restriction digestions of λmA14 and λmA36 produced two hybridising fragments, both of which must contain actin DNA, and therefore these sites occurred within the actin coding region of both clones. This was consistent with the position of these sites within λmA19, as indicated in Figure 2.3.

The position of the sites within and surrounding the actin-like regions was determined with more accuracy by double restriction digestion followed by hybridisation to a ^{32}P-labelled actin probe. The double restriction digestions were initially performed using BglII for the first digestion, as a BglII site was known to occur within the actin-coding region and its position could be predicted from the nucleotide sequence of λmA19. Figure 3.4 is an example of such double digestions of λmA14 and λmA36 hybridised against a ^{32}P-labelled PstI fragment of the skeletal muscle cDNA clone, pmS3 (Figure 2.2). The fragments which hybridised to the probe would identify the first restriction site 3' to the BglII site for any given restriction enzyme provided a site occurred before the second BglII site, 4.7kb to the right. Figure 3.4 illustrates the results of BglII double digestions with EcoRI, SmaI and PstI of λmA14 and λmA36. All gave rise to a 4.7kb hybridising fragment (The 4.7kb BglII fragment), indicating that none of these restriction enzymes has a site within 4.7kb of the BglII site in either clone. BglII-XbaI double digestion of λmA14 and λmA36 produced a 850bp hybridising fragment, and thus indicated that a XbaI site occurred 850bp to the right of the BglII site in both

Figure 3.4 Example of products of BglII double digestion of λmA14 and λmA36 hybridised against ^{32}P-labelled actin probe

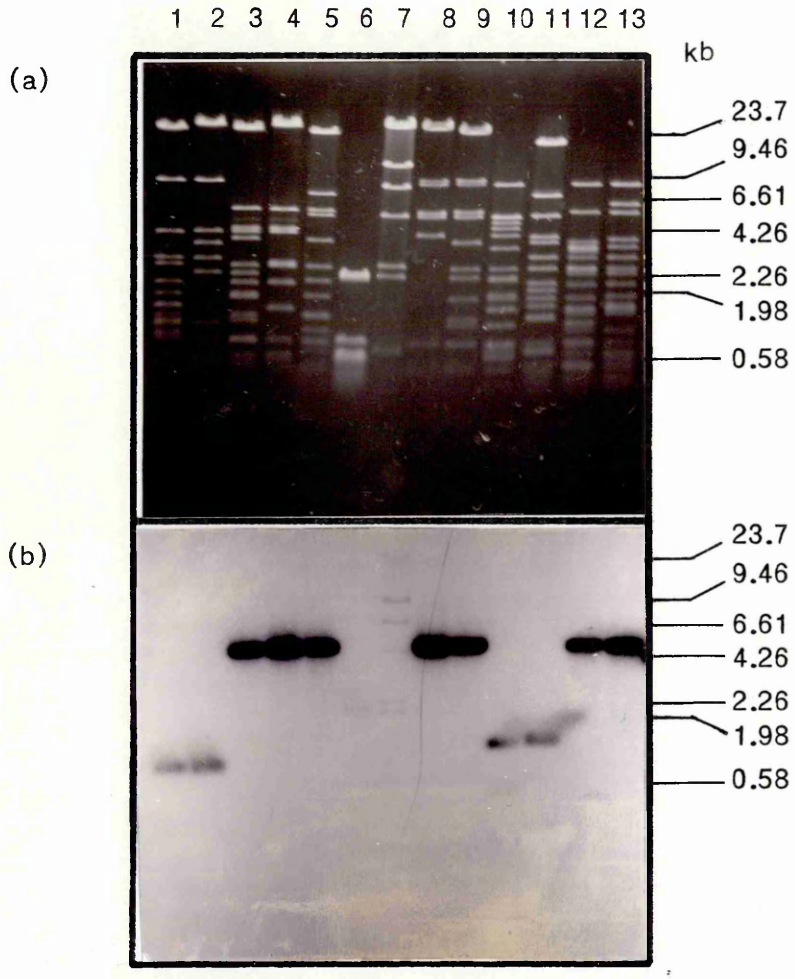
λmA14 and λmA36 were analysed as previously described in Figure 3.3, BglII being the primary restriction enzyme in all cases.

(a) Photograph of the stained gel

(b) Autoradiograph of the nitrocellulose

The fragment(s) which hybridised to the actin probe are indicated below

Lane	DNA	Restriction enzyme 1	Restriction enzyme 2	Hybridised fragment(s) (kb)
1	λmA14	BglII	XbaI	0.85
2	λmA36	BglII	XbaI	0.85
3	λmA14	BglII	EcoRI	4.7
4	λmA36	BglII	EcoRI	4.7
5	λmA14	BglII	SmaI	4.7
6	pBR322	BglI	BamHI	-
7	λcI_{857}	HindIII		-
8	λmA14	BglII		4.7
9	λmA36	BglII		4.7
10	λmA14	BglII	PvuII	1.4 and 0.35
11	λmA36	BglII	PvuII	1.5
12	λmA14	BglII	PstI	4.7
13	λmA36	BglII	PstI	4.7


1 2 3 4 5 6 7 8 9 10 11 12 13
kb
(a)
23.7
9.46
6.61
4.26
2.26
1.98
0.58
(b)
23.7
9.46
6.61
4.26
2.26
1.98
0.58

genomic clones. This result was consistent with the BglII and XbaI sites in λmA14 and λmA36 being in the same relative position as found in λmA19, were they are shown to be 850bp apart (Figure 2.3).

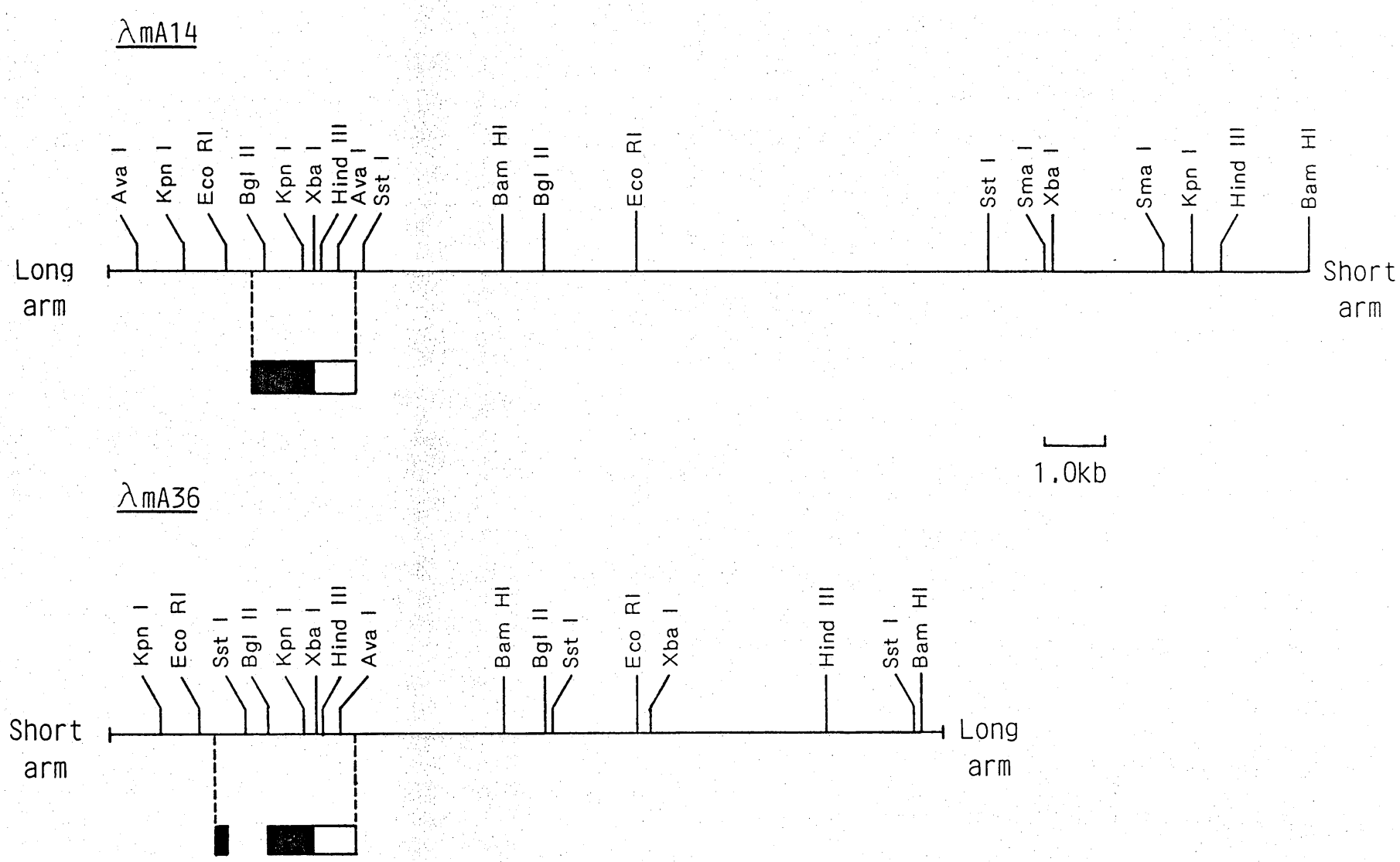
In order to map sites 5' to the XbaI site, double digestions were performed with XbaI and these were hybridised against a ^{32}P-labelled actin coding probe. The XbaI-EcoRI double digestion of λmA14 and λmA36, respectively produced different sized labelled hybridising fragments of lengths 1.5 and 2.0kb. As discussed in the Introduction, the electron micrographs show that 540bp of extra DNA interrupts the actin-like coding region of λmA36, but not of λmA14. This difference could therefore allow these two EcoRi sites in λmA14 and λmA36 to be equivalent despite their difference in distance from the XbaI site. (A similar arguement applies to the results of the single restriction digestion with EcoRI or KpnI hybridised against a ^{32}P-labelled actin probe, see Table 3.1). As the BglII-XbaI double digestions of λmA14 and λmA36 had given a similar 850bp fragment, this indicated that the extra DNA in λmA36 did not occur between BglII and the XbaI site.

Double digestions were performed with EcoRI, so as to use as a reference point the EcoRI sites located 5' to the actin regions in λmA14 and λmA36. It was hoped in this way to extend further the mapping 3' to the actin regions, as the next EcoRI site occurred 7.0 and 7.5kb, respectively to the right, 1.7kb beyond the 5' flanking BglII site. However these blots did not provide any new information and only confirmed the positions of the restriction sites previously mapped by the BglII double digestions.

Figure 3.5 shows the partial restriction maps of λmA14 and λmA36 revised using the hybridisation results from the single and double digestions. This shows that there are an increasing number of sites at similar positions

Figure 3.5 Partial restriction maps of λmA14 and λmA36 (version II)

The partial restriction maps of λmA14 and λmA36 were revised using the results from the single and double digestions followed by hybridisation to an actin probe. The positions of the actin pseudogene regions are predicted from electron microscopy and are indicated as in Figure 3.2.


λmA14
Ava I
Kpn I
Eco RI
Bgl II
Kpn I
Xba I
Hind III
Ava I
Sst I
Bam HI
Bgl II
Eco RI
Sst I
Sma I
Xba I
Sma I
Kpn I
Hind III
Bam HI
Long arm
Short arm
1.0kb
λmA36
Kpn I
Eco RI
Sst I
Bgl II
Kpn I
Xba I
Hind III
Ava I
Bam HI
Bgl II
Sst I
Eco RI
Xba I
Hind III
Sst I
Bam HI
Short arm
Long arm

in both clones.

In order to obtain comprehensive restriction maps over the whole of the inserts of λmA14 and λmA36 a different method was required. Rather than the tedious and difficult traditional method of double restriction digestions, the partial mapping technique of Rackwitz *et al.*, (1984) was employed, (section 2.5). Figure 3.6 shows the results of such an analysis for λmA36 partially digested with several enzymes. The samples of partially digested λmA36 were mixed with the ^{32}P-labelled oligonucleotide complementary to the cohesive end of the short arm of lambda and then subjected to electrophoresis. The autoradiograph of the gel visualises those partial digestion products which contain this cohesive end, the length of the fragment indicating the distance of the restriction site, from the end of the short arm. The restriction map for a particular enzyme can be read off the autoradiograph in a manner analogous to reading a sequencing gel. For example, the results of partial digestion of λmA36 with EcoRI produced four labelled fragments of lengths, 3.5, 10.5, 18.0 and >23.0kb. The 3.5kb fragment was generated by cleavage at the EcoRI site in the short arm (9kb) of the vector, 3.5kb from its cohesive end. The 10.5kb fragment was generated by cleavage at a EcoRI site which occurred 1.5kb into the insert (to the right of its extremity as present in Figure 3.7), and the 18.0kb fragment locates the next EcoRI site, 9kb into the insert. The largest fragment represented undigested λmA36. Table 3.2 summaries the lengths of the labelled products generated when λmA36 is partially digested with a variety of restriction endonucleases.

A second example illustrates this method in a more difficult case, that of λmA14 digested with AvaI, where the results of complete digestion, (Figure

Figure 3.6 Example of λmA36 mapped by the partial digestion technique

λmA36 was subjected to partial digestion mapping (Rackwitz *et al.*, 1984) as described in section 2.5. The autoradiograph of the dried gel is shown.The sizes of the labelled fragments for each restriction endonucleases are listed in Table 3.2.

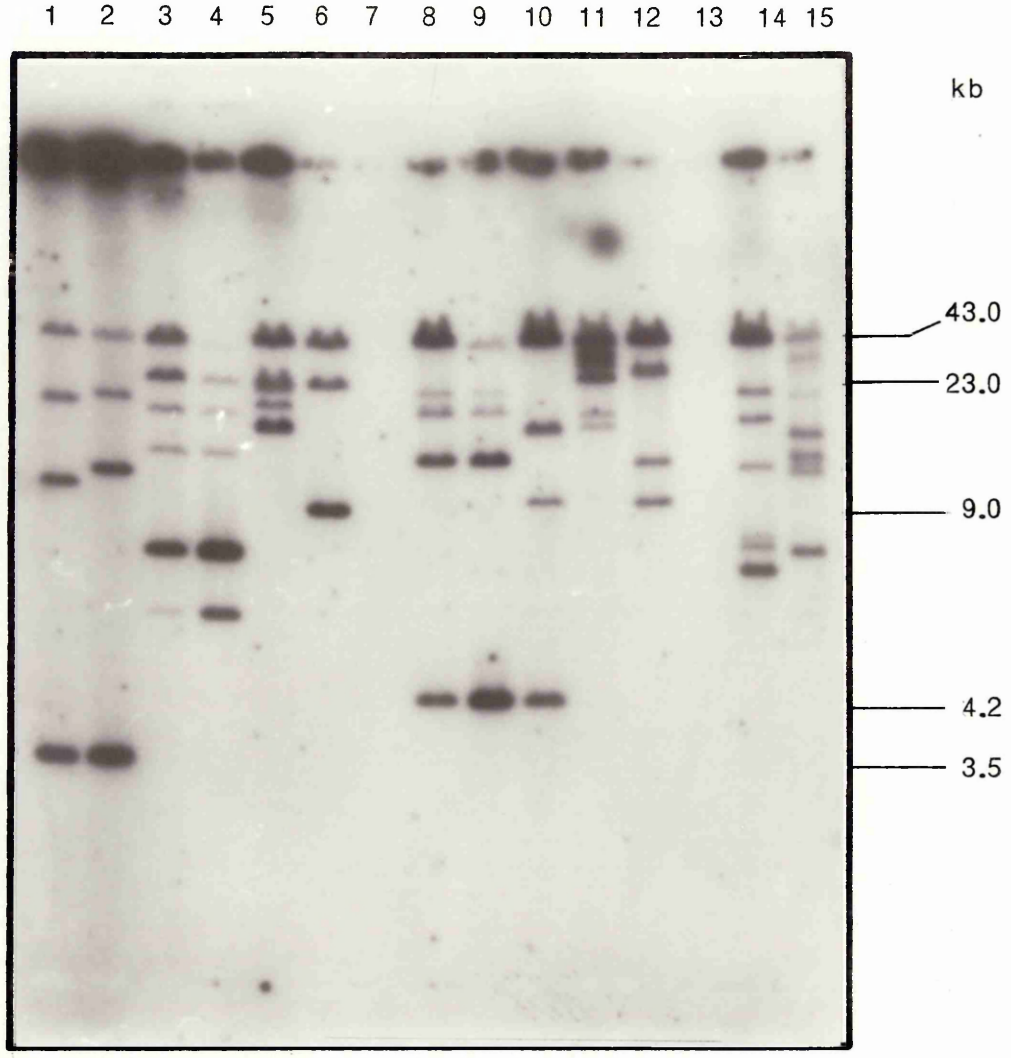

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
kb
43.0
23.0
9.0
4.2
3.5

Table 3.2 Lengths of labelled fragments produced by partial digestion of λmA14 and λmA36

Lane	DNA	Restriction enzyme	Partial digestion fragments (kb)
1	λmA36	EcoRI	3.5 10.5 18.0
2	λ1059	EcoRI	3.5 12.0 20.0
3/4	λmA36	AvaI	5.6 7.4 13.0 17.5 23.0
5	λmA36	BamHI	15.5 16.0 19.0 21.5 23.0
6	λ1059	BamHI	9.0 23.0 43.0
7	λcI_{857}	HindIII	-
8/9	λmA36	HindIII	4.2 12.5 19.0 20.0 22.5
10	λ1059	HindIII	4.2 9.5 15.5
11	λmA36	PvuII	16.0 17.5
12	λmA36	KpnI	10.0 12.5
14	λmA36	BglII	7.0 7.0-9.0 (numerous) 11.5 16.5 20.5
15	λmA14	AvaI	7.4 11.0 11.5 12.0 14.5 15.0 20.0

3.1) are too complex to be interpreted on their own. Figure 3.6 shows the labelled products of the partial digestion of λmA14 with AvaI. These were fragments of lengths 7.4, 11.0, 11.5, 12.0, 14.5, 15.0, 20.0 and >23.0kb. AvaI cleaves the short arm of lambda twice (Karn *et al.*, 1980), however the partial digestion resulted in cleavage at only one of these sites to generate 7.4kb fragment. The 11.0kb fragment was generated by cleavage at an AvaI site located 2.0kb into the insert (to the left of its extremity as presented in Figure 3.7), and subsequent fragments were produced by cleavage at sites 2.5, 3.0, 6.0, 11.0 and >14.0kb into the insert. Although this technique allowed unambiguous ordering of the sites, the distance between them could only be calculated from differences between fragments, and some of these differences were relatively small compared to the sizes of the fragment measured. The value of the sizes of the smaller fragments resulting from complete digestion (Figure 3.1) were therefore used to refine the map.

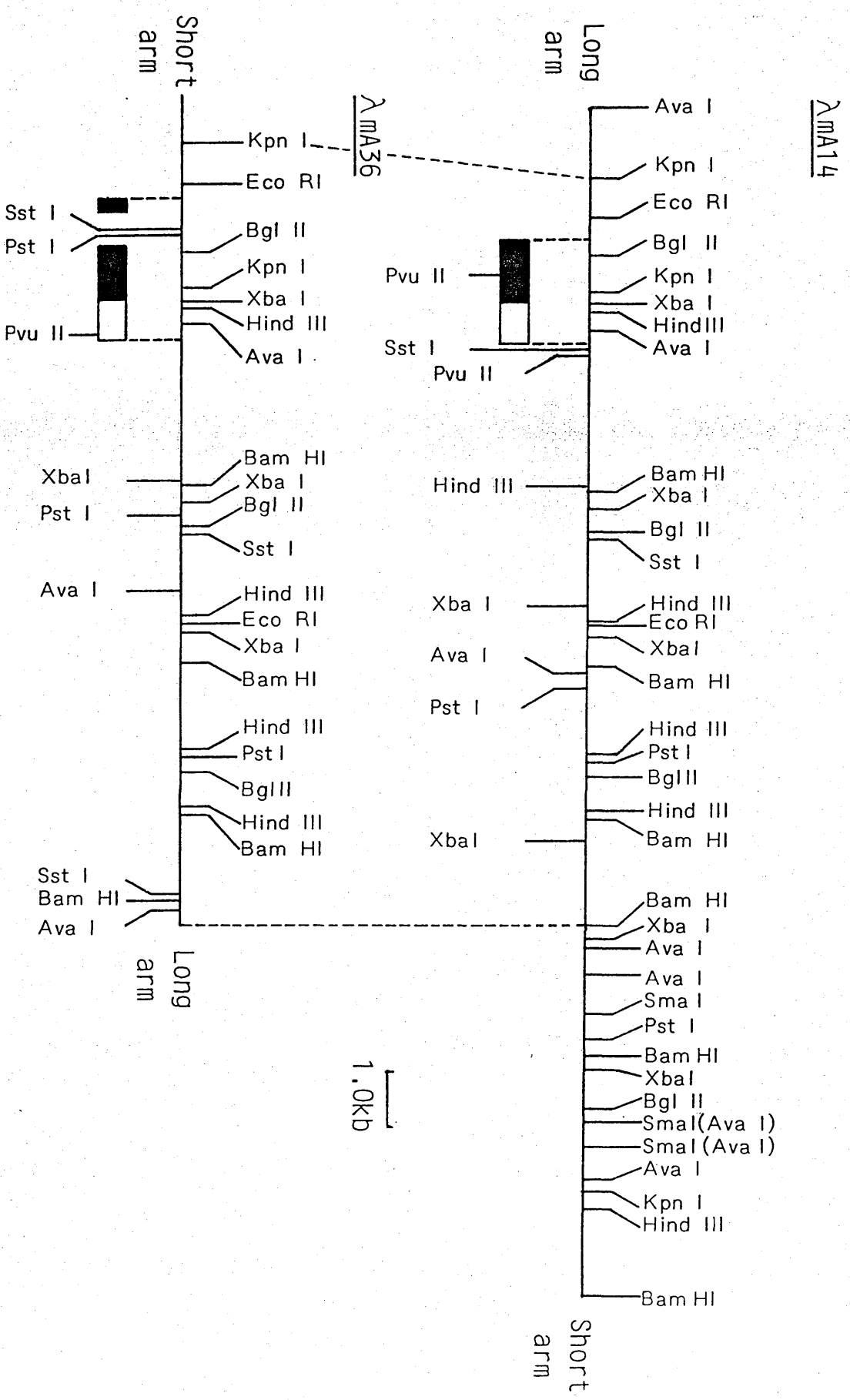
The revised restriction maps using these data are shown in Figure 3.7. They still contain ambiguites relating to the order of closely adjacent sites and the possibility of sites missed because of the partial digestion. However it is evident that they allow comparison of a large number of sites over the whole of the inserts within λmA14 and λmA36.

3.1.2 Derivation and analysis of subclones of λmA14 and λmA36

The maps in Figure 3.7, formed the basis for determining a strategy to subclone much of λmA14 and λmA36. As well as being a necessary preliminary to the sequence analysis described in sections 3.1.4 and 3.2, the generation of subclones was important for the comparison of λmA14 and λmA36, as restriction mapping of the small inserts of such subclones allowed

Figure 3.7 Partial restriction maps of λmA14 and λmA36 (versionIII)

The partial restriction maps of λmA14 and λmA36 were revised using the results of the partial digestion technique. The positions of the actin pseudogene regions predicted from electron microscopy are as indicated in Figure 3.2.


Short arm
λmA36
Kpn I
Eco RI
Sst I
Pst I
Bgl II
Kpn I
Xba I
Hind III
Pvu II
Ava I
Bam HI
XbaI
Xba I
Pst I
Bgl II
Sst I
Ava I
Hind III
Eco RI
Xba I
Bam HI
Hind III
Pst I
BglII
Hind III
Bam HI
Sst I
Bam HI
Ava I
Long arm
Long arm
λmA14
Ava I
Kpn I
Eco RI
Bgl II
Pvu II
Kpn I
Xba I
HindIII
Sst I
Ava I
Pvu II
Hind III
Bam HI
Xba I
Bgl II
Sst I
Xba I
Hind III
Eco RI
XbaI
Ava I
Bam HI
Pst I
Hind III
Pst I
BglII
Hind III
XbaI
Bam HI
Bam HI
Xba I
Ava I
Ava I
Sma I
Pst I
Bam HI
XbaI
Bgl II
SmaI(Ava I)
SmaI(Ava I)
Ava I
Kpn I
Hind III
Bam HI
Short arm
1.0kb

much more precise correlation of sites, apparently at similar positions in Figure 3.7.

The general methods by which the subclones were constructed and identified are described in detail in section 2.4. The relationship of the subclones to the parent genomic clones, λmA14 and λmA36, is shown in Figure 3.8, and detailed restriction maps of individual subclones are shown in Figure 3.9 to 3.14. The basis for the identification and positioning of each of these subclones shown in Figure 3.7 was as follows.

λmA14 and λmA36 subclones, 14KK1 (2.0kb) and 36KK1 (2.5kb), were identified using a ^{32}P-labelled PstI fragment from pmS3 (Figure 2.2, predominately γ-actin coding region), and their identities confirmed by the presence of EcoRI and BglII sites (Figure 3.9) in positions consistent with the overall restriction maps (Figure 3.7).

λmA14 HindIII subclone, 14HH1 (3.0kb) and λmA36 XbaI subclone 36XX1 (3.1kb), were identified using a ^{32}P-labelled PstI-XbaI fragment from the subclone MγA-ψ1, a subclone of λmA19 (Figure 2.3, which contains the γ-actin 3'non-coding region). Their identities were confirmed by the presence of the AvaI site in positions consistent with the overall restriction maps (Figure 3.7). The presence of a SstI site in 14HH1 allowed two further subclones to be derived from this, and facilitated subsequent sequencing. The restriction maps of 14HH1 and 36XX1 are shown in Figure 3.10.

The λmA14 HindIII subclones, 14HH2 (2.3kb), 14HH3 (1.0kb) and 14HH4 (2.3kb) were identified by relating their restriction maps to the overall maps (Figure 3.7). The position and orientation of the subclone 14HH2 was deduced from the presence and location of a SstI site (Figure 3.11), not present in either 14HH3 and 14HH4. The position and orientation of the subclone 14HH4 was deduced from the presence and location of the BglII site (Figure 3.11).

Figure 3.8 Relationship of subclones to the parent genomic clones λmA14 and λmA36

The general methods by which the subclones were constructed and identified are described in section 2.4 and in the text of the Result chapter. The position of the actin pseudogene regions predicted from electron microscopy are as indicated in Figure 3.2.

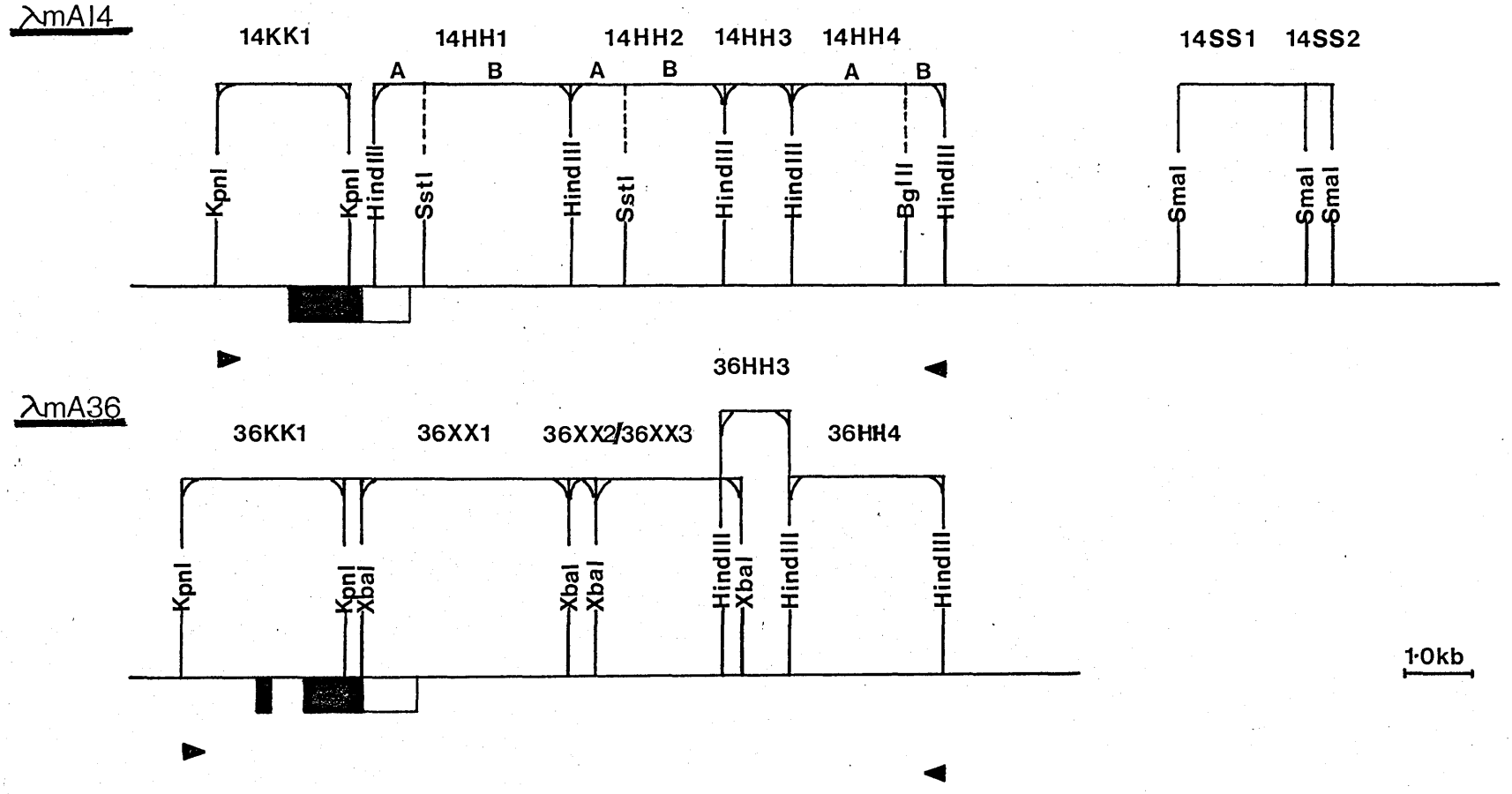

λmA14
14KK1
14HH1
14HH2
14HH3
14HH4
14SS1
14SS2
A
B
A
B
A
B
KpnI
KpnI
HindIII
SstI
HindIII
SstI
HindIII
HindIII
BglII
HindIII
SmaI
SmaI
SmaI
36HH3
λmA36
36KK1
36XX1
36XX2/36XX3
36HH4
KpnI
KpnI
XbaI
XbaI
XbaI
HindIII
XbaI
HindIII
HindIII
1·0kb

Figure 3.9 Partial restriction maps of λmA14 and λmA36 KpnI subclones 14KK1 and 36KK1

The maps show selected restriction sites in the inserts of clones 14KK1 and 36KK1 in pUC18. Mapping was by single and double digestion with the endonucleases for the sites indicated together with the endonucleases that cleave once in the polylinker.

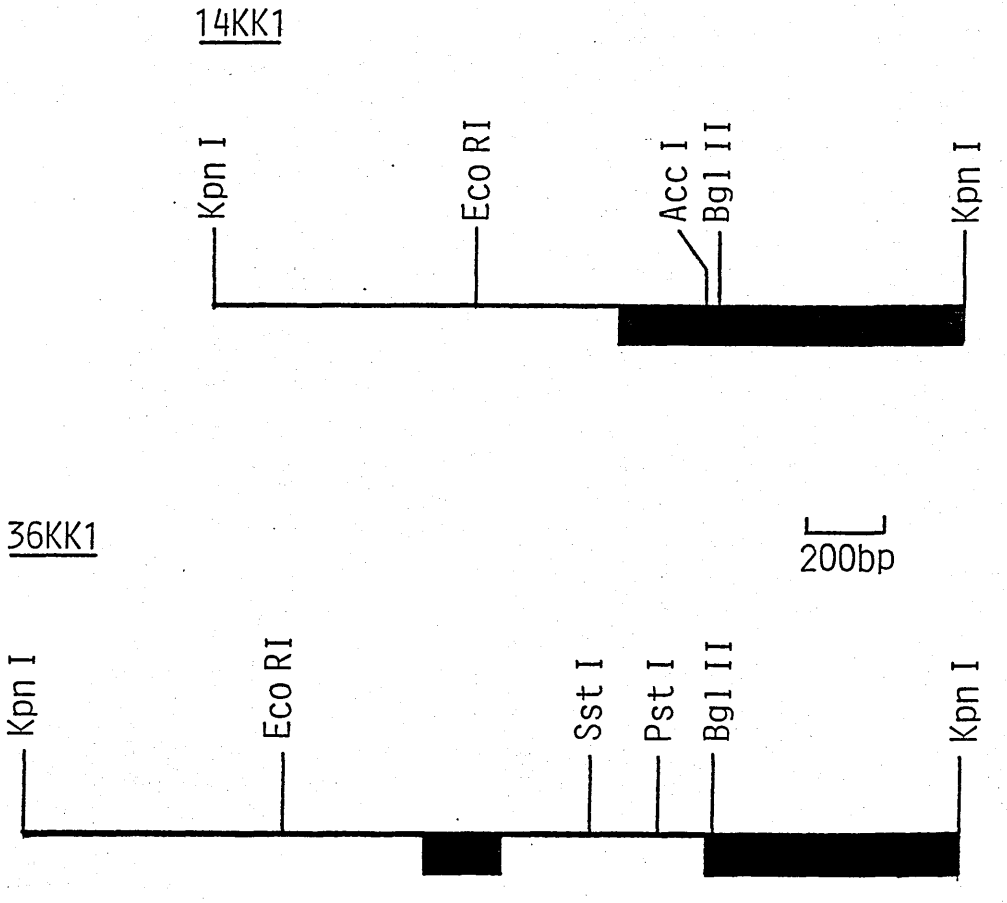
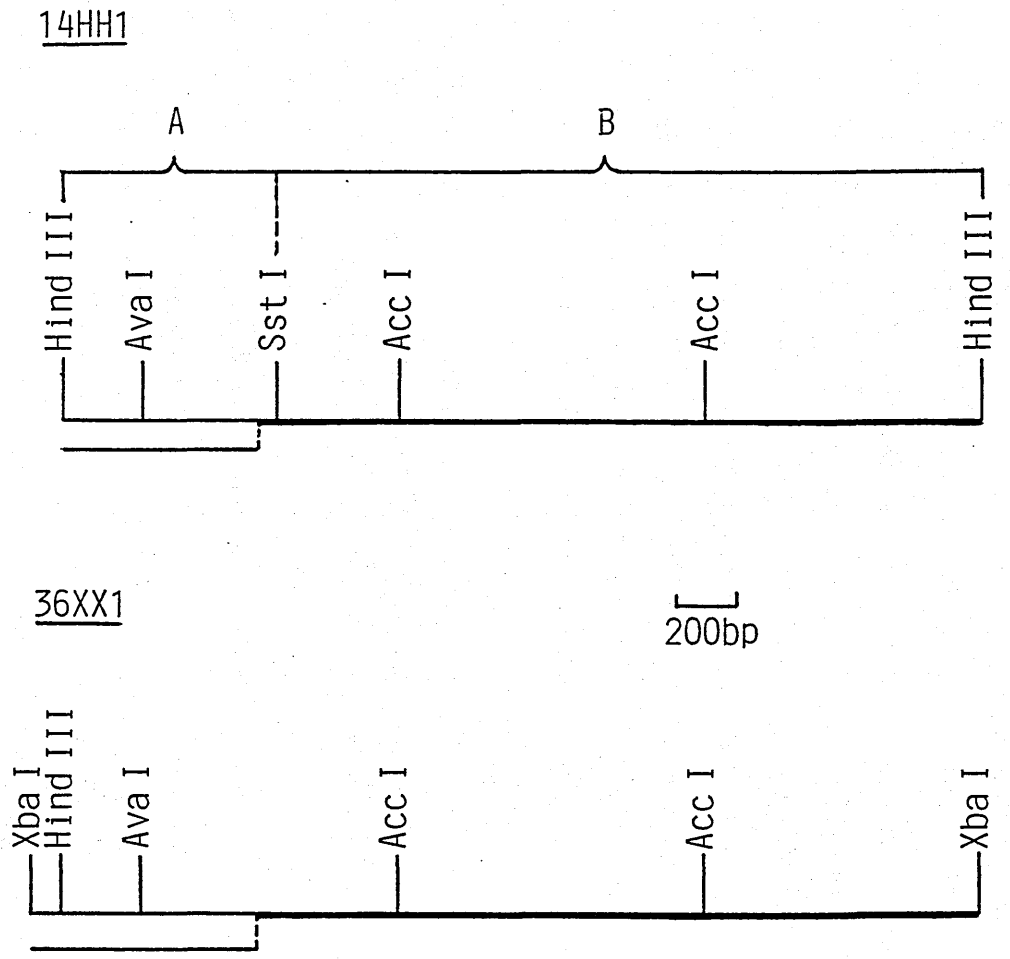


Figure 3.10 Partial restriction maps of λmA14 HindIII subclone 14HH1 and λmA36 XbaI subclone 36XX1

The maps show selected restriction sites in the inserts of clones 14HH1 and 36XX1 in pUC18. Mapping was as described in Figure 3,9. Subclones 14HH1A and 14HH1B were produced by digesting both orientations of 14HH1 with SstI followed by religation.



Subsequent sequencing of the subclone (described in section 3.2.2) proved this orientation to be correct. The 1.0kb HindIII fragment, subcloned into 14HH3 has previously been tentatively assigned a location to the right of the fragments subcloned into 14HH2 and 14HH4 (see Figure 3.7). However mapping of 14HH3 indicated that it contained several restriction sites which clearly positioned it between 14HH2 and 14HH4. The orientation of 14HH3, although based on restriction mapping must still be considered uncertain, because the proximity of the sites made unambiguous ordering of the partial digestion fragment (Figure 3.6) difficult. Mapping of the λmA36 subclones 36XX3 and 36HH3 (see below) shows that there are sites in the corresponding region of theses clones in a similar order to that proposed for 14HH3. (The order of restriction enzyme sites could be more precisely determined in these subclones of λmA36 because they overlap). Therefore the designation of 14HH3, shown in Figure 3.11, is most likely to be correct. The subclones 14HH2 and 14HH4 were further subcloned using, respectively, the internal SstI and BglII sites. The restriction maps of 14HH2, 14HH3 and 14HH4 are shown in Figure 3.11.

The λmA36 HindIII subclones 36HH3 (1.0kb) and 36HH4 (2.3kb), were identified as follows. The position and orientation of the subclone 36HH4 was deduced as for 14HH4, from the presence and location of the BglII site. The position of 36HH3 was determined as described for 14HH3. The orientation of 36HH3 was determined by the presence of EcoRI and XbaI, 100 and 300bp respectively from the left-hand HindIII site. The sites fell within a region contained in the subclone 36XX3 which overlapped 36HHβ on the left-hand side by 300bp. The restriction maps of 36HH3 and 36HH4 are shown in Figure 3.12.

The λmA36 XbaI subclones, 36XX2 (400bp) and 36XX3 (2.3kb), were

Figure 3.11 Partial restriction maps of λmA14 HindIII subclones 14HH2, 14HH3 and 14HH4

The maps show selected restriction sites in the inserts of clones 14HH2, 14HH3 and 14HH4. Mapping was as described in Figure 3.9. Subclones 14HH2A and 14HH2B were produced by digesting both orientations of 14HH2 with SstI followed by religation. Subclones 14HH4A and 14HH4B were produced by digesting both orientations of 14HH4 with BglII and BamHI followed by religation.

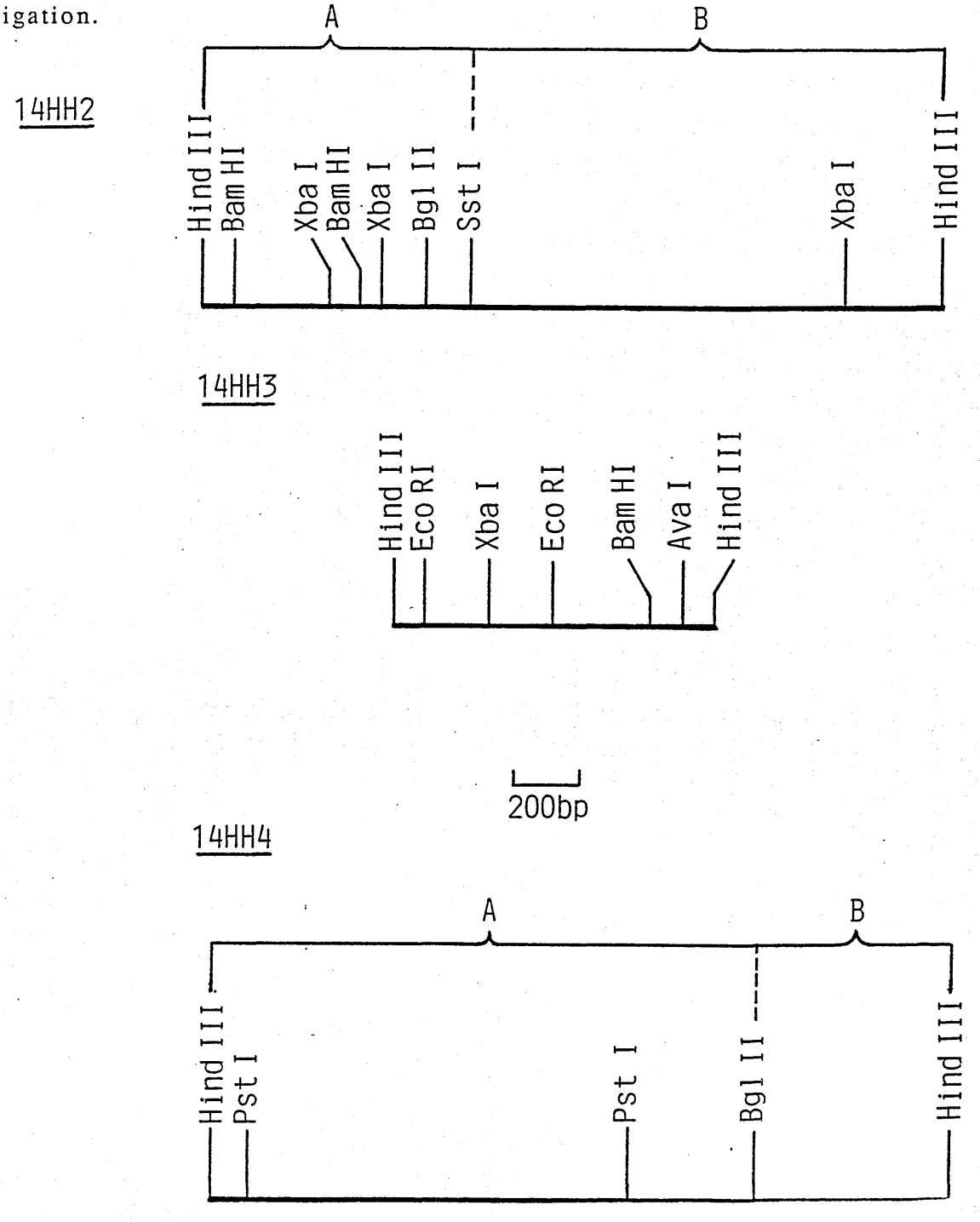
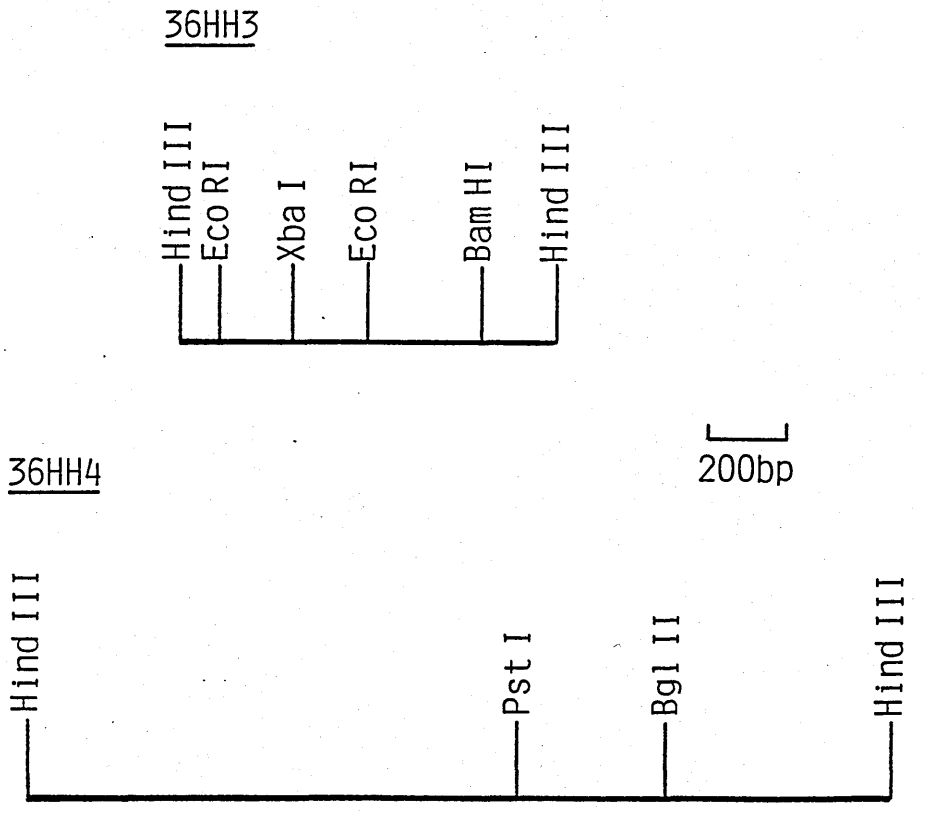


Figure 3.12 Partial restriction maps of λmA36 HindIII subclones 36HH3 and 36HH4

The maps show selected restriction sites in the inserts of clones 36HH3 and 36HH4 in pUC18. Mapping was as described in Figure 3.9.



identified as follows. Subclone 36XX3 was found to overlap subclone 36HH3 (as described above) and was therefore positioned and oriented by the presence of the EcoRI and HindIII sites near the 3' XbaI site. The subclone 36XX2 was located to the left of 36XX3 as the 400bp XbaI fragment within 36XX2 was required to position the sites within 36XX3 at the correct predetermined distance from the actin region. Further confirmation that the location of 36XX2 was correct, was that 36XX3 was predicted to have two BamHI sites within 500bp of the 5' XbaI site, however there was only one, the second was located in the subclone to the left, 36XX2. Restriction maps of 36XX2 and 36XX3 are shown in Figure 3.13.

The λmA14 subclones designate by 'SmaI' were actually produced using the isoschizomer, XcyI, which cleaves to produce 5' protruding ends rather than blunt ends produced by cleavage with SmaI, and therefore increases the effeciency of ligation. The subclones, 14SS1 (1.9kb) and 14SS2 (400bp) shown in Figure 3.8 were from a region of λmA14 which fell outwith the genomic region of λmA36 and were not used for comparitive purposes. The 14SS1 subclone included the region a_R (Figure 1.8) and was identified using a probe containing a_L, as described in detail in section 3.2.1, below. The presence of the three PstI sites within 14SS1 allowed three further subclones to be derived from the two internal PstI fragments, 970 and 330bp in length, and the PstI to 3' SmaI fragment of 470bp, and these facilitated subsequent sequencing. The maps are presented here for consistency (Figure 3.14).

The restriction maps of λmA14 and λmA36 were revised in the light of the mapping of the subclones and the final version of these maps are shown in Figure 3.15. The main revisions were the placing of subclones 14HH3, 36HH3 and 36XX2 (see above), the revision of the order of clustered restriction

Figure 3.13 Partial restriction maps of λmA36 XbaI subclones 36XX2 and 36XX3

The maps show selected restriction sites in the inserts of clones 36XX2 and 36XX3 in pUC18. Mapping was as described in Figure 3.9.

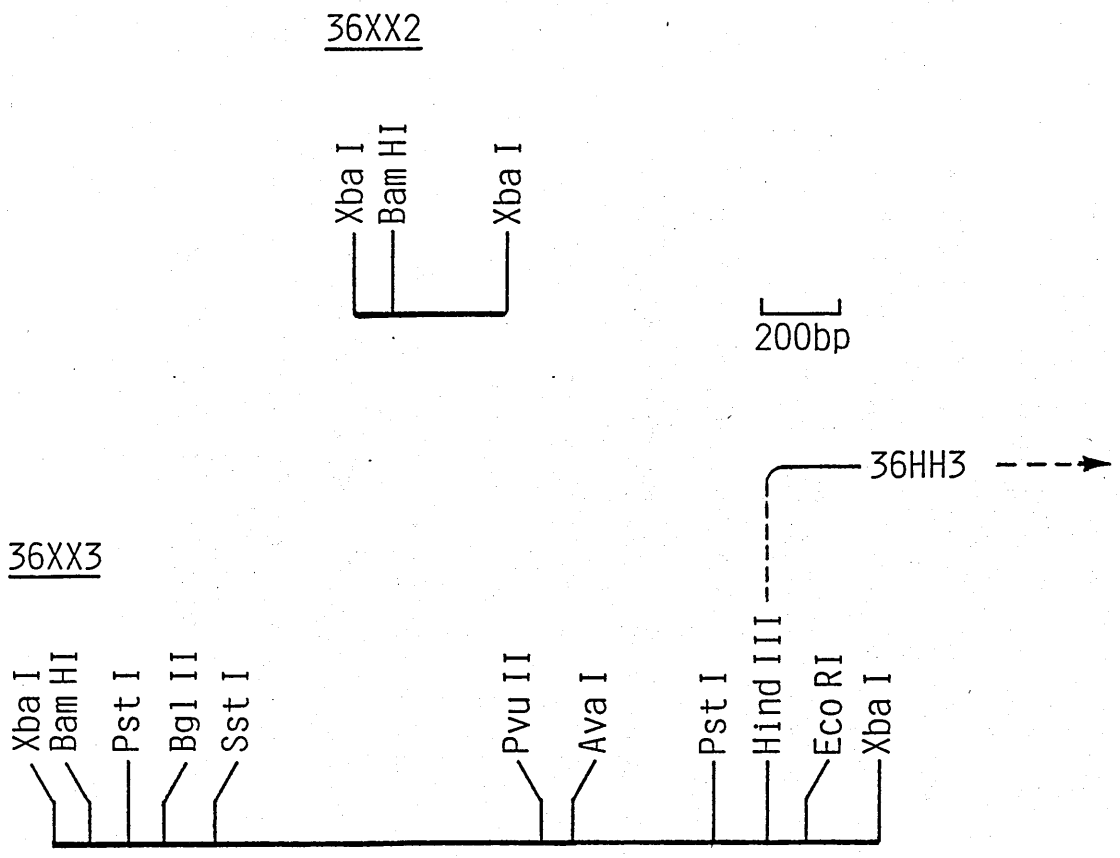
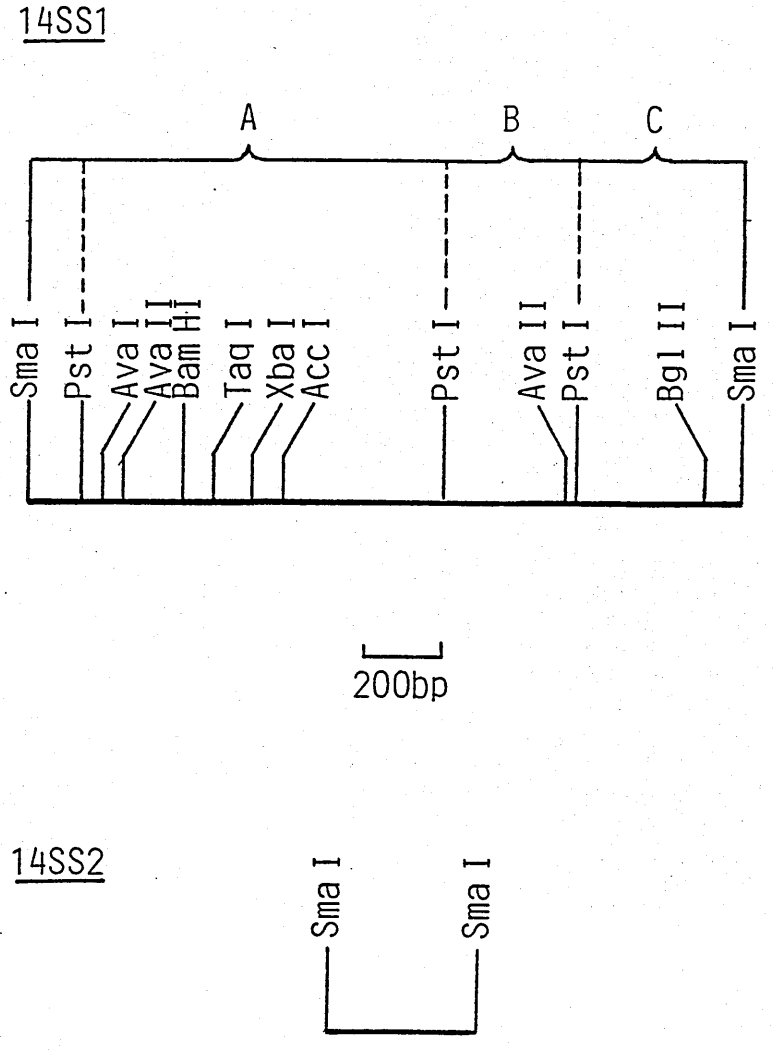


Figure 3.14 Partial restriction maps of λmA14 'SmaI' subclones 14SS1 and 14SS2

The maps show selected restriction sites in the inserts of clones 14SS1 and 14SS2 in pUC18. Mapping was as described in Figure 3.9. The subclones 14SS1A and 14SS1B were produced by resubcloning the two internal PstI fragments of lengths of 0.97 and 0.33kb. Subclone 14SS1C was produced by digesting 14SS1 with PstI followed by religation.



sites, for example, within the 1.0kb HindIII subclones, and the detection of a couple of unknown sites previously hidden by the close proximity of similar sites, for example, EcoRI sites in the HH3 subclones.

Comparison of the detailed restriction maps of λmA14 and λmA36 indicated that their similarity appeared to extend over 11.0kb (with respect to λmA14) from the KpnI site in 14KK1 to the BamHI site 100bp beyond the 3' HindIII site of 14HH4.

3.1.3 Cross-hybridisation between λmA14 and λmA36

The subcloning of much of λmA14 and λma36 allowed confirmation of their relatedness by cross-hybridising fragments from λmA14 against λmA36. The λmA14 restriction fragments were isolated from the subclone 14HH1B (Figure 3.16) which contains the DNA of the left-hand arm of the foldback structure in this clone.

Figure 3.17 shows the ^{32}P-labelled SstI-AccI restriction fragment from the subclone 14HH1B, hybridised against λmA36 digested with various restriction enzymes. The ^{32}P-labelled probe hybridised to a 5.3kb HindIII restriction fragment of λmA36. This fragment is located from 3.6 to 8.9kb from the left extremity of the insert in λmA36 (Figure 3.15) and hence includes the region corresponding to the position of the SstI-AccI fragment of λmA14. This indicated that λmA36 contained DNA homologous to that in the SstI-AccI fragment, and was consistent with it being at a similar location with respect to the actin pseudogene.

Figure 3.18 shows the 1.0kb AccI restriction fragment from 14HH1B

Figure 3.15 Final partial restriction maps of λmA14 and λmA36 (version IV)

The partial restriction maps of λmA14 and λmA36 were revised using the results from the subcloning. The positions of the actin pseudogene regions predicted from electron microscopy was as indicated in Figure 3.2. Within the region enclosed by the dashed line, restriction sites which are similarly positioned in λmA14 and λmA36 are above the line and those which differ below. Outwith the enclosed region, all the restriction sites are above the line. The maps are complete for BglII, PvuII and PstI only in the regions which have been subcloned. Sites for AccI have only been mapped in specific subclones.

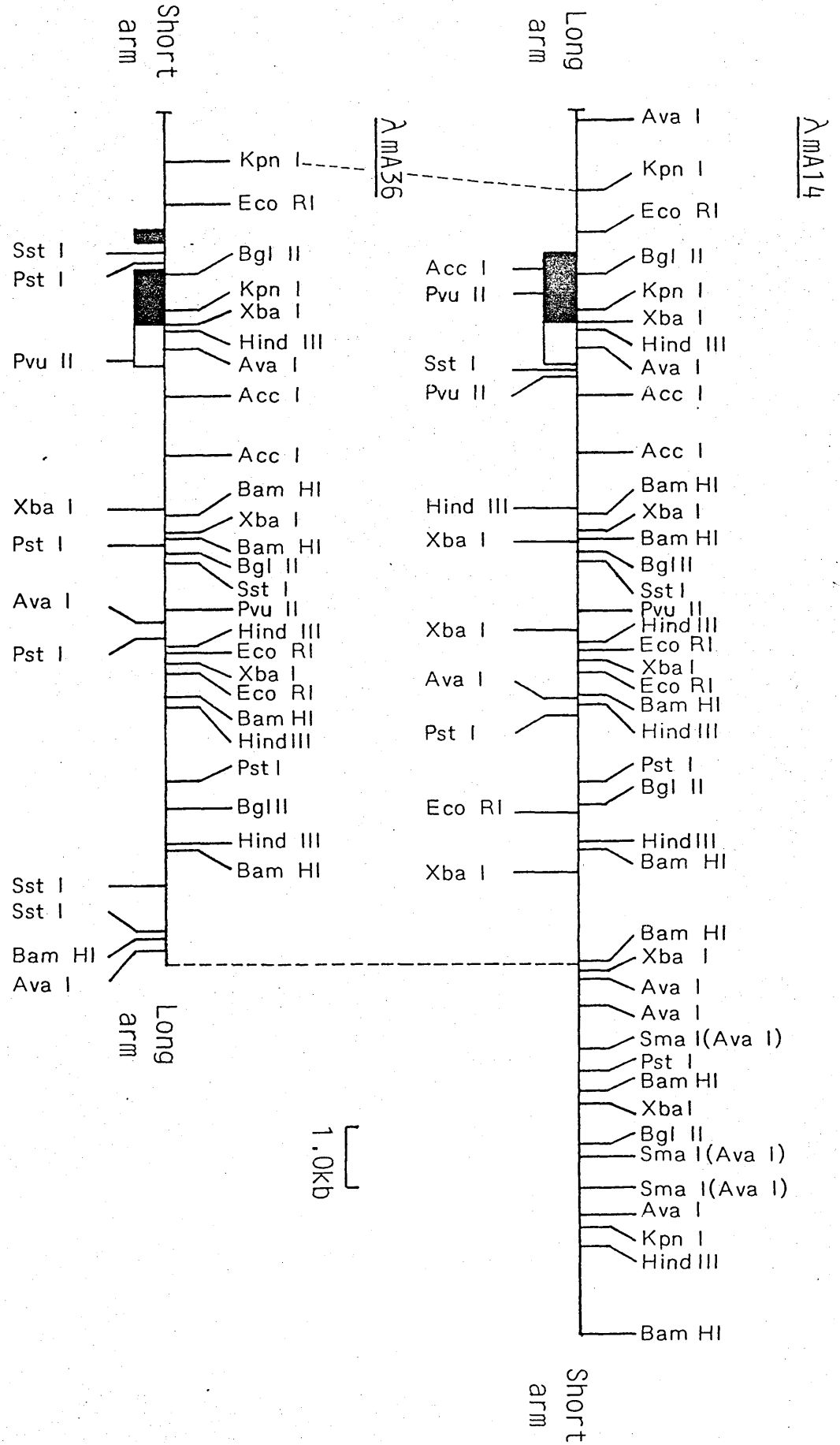

Short arm
λmA36
Kpn I
Eco RI
Sst I
Pst I
Bgl II
Kpn I
Xba I
Hind III
Pvu II
Ava I
Acc I
Acc I
Bam HI
Xba I
Xba I
Pst I
Bam HI
Bgl II
Sst I
Ava I
Pvu II
Hind III
Pst I
Eco RI
Xba I
Eco RI
Bam HI
Hind III
Pst I
BglII
Hind III
Bam HI
Sst I
Sst I
Bam HI
Ava I
Long arm
1.0kb
Long arm
λmA14
Ava I
Kpn I
Eco RI
Acc I
Bgl II
Pvu II
Kpn I
Xba I
Sst I
Hind III
Pvu II
Ava I
Acc I
Acc I
Hind III
Bam HI
Xba I
Xba I
Bam HI
BglII
Sst I
Pvu II
Xba I
Hind III
Eco RI
Xba I
Ava I
Eco RI
Bam HI
Pst I
Hind III
Pst I
Bgl II
Eco RI
Hind III
Bam HI
Xba I
Bam HI
Xba I
Ava I
Ava I
Sma I(Ava I)
Pst I
Bam HI
Xba I
Bgl II
Sma I(Ava I)
Sma I(Ava I)
Ava I
Kpn I
Hind III
Bam HI
Short arm

Figure 3.16 Location of DNA probes isolated from subclone 14HH1B, used to hybridise to digested λmA36

The λmA14 subclone 14HH1B was digested with SstI, AccI and HindIII (section 2.2.8) and three fragments (i), (ii) and (iii) isolated (section 2.2.13). The DNA probes were :

(i) SstI-AccI 400bp fragment

(ii) AccI 1.0kb fragment

(iii) AccI-HindIII 900bp fragment

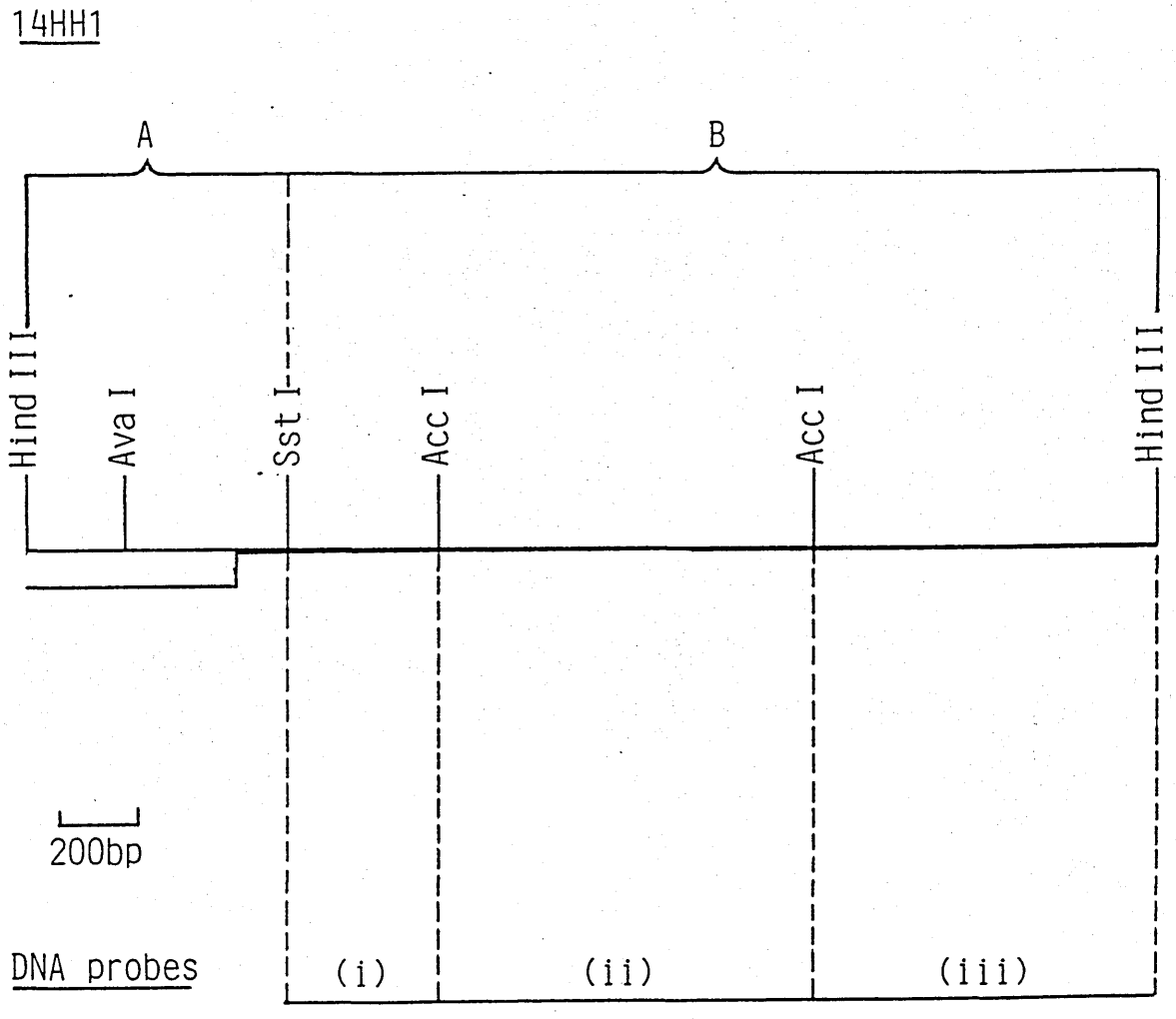


Figure 3.17 Hybridisation of SstI-AccI fragment from subclone 14HH1B, against digested λmA36

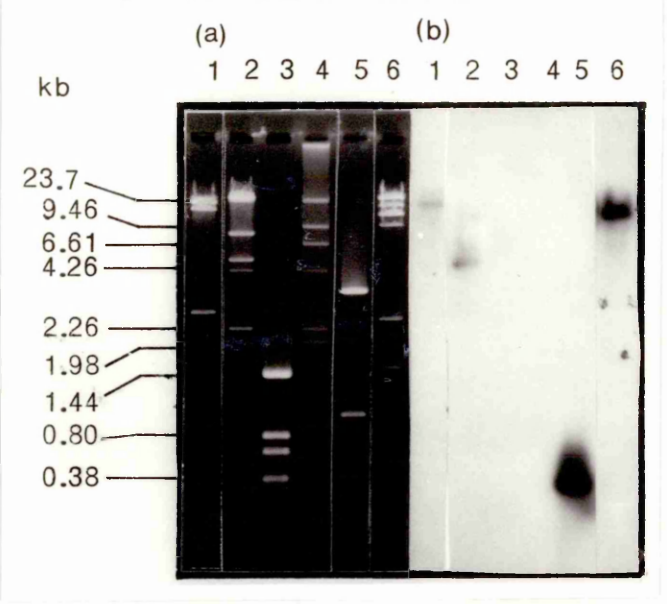
λmA36 was digested with the restriction endonucleases indicated (section 2.2.8) and subjected to electrophoresis through a 1% agarose gel (section 2.2.10). The DNA was transferred to the nitrocellulose (section 2.2.15) and hybridised against the ^{32}P-labelled SstI-AccI fragment (section 2.2.17) from the subclone 14HH1B (Figure 3.16).

(a) Photograph of the stained gel (lanes 1-6)

(b) Autoradiograph of the nitrocellulose (lanes 1-6)

The fragment(s) which hybridised to the ^{32}P-labelled DNA probe are indicated below :

Lane	DNA	Restriction enzyme	Hybridised fragment(s) (kb)
1	λmA36	BamHI	15.7
2	λmA36	HindIII	5.3
3	λcI_{857}	HindIII	-
4	pmS4	TaqI	-
5	14HH1B	SstI/AccI	0·4
6	λmA36	KpnI	12.5


(a)
(b)
1 2 3 4 5 6 1 2 3 4 5 6
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

hybridised against λmA36 digested with various restriction enzymes. The AccI fragment hybridised to both a 5.3 and 2.3kb HindIII restriction fragment of λmA36. The 5.3kb fragment of λmA36 includes the region corresponding to the AccI fragment of λmA14 and hence the result is consistent with homologous DNA at equivalent locations in the two clones. The hybridisation to the 2.3kb HindIII fragment indicated that DNA homologous to the probe also occurred in a second location within λmA36. This is because the AccI fragment includes part of the b_L region of the stem, which has a complementary region b_R, in λmA14 (see below) and, if λmA36 is homologous, it would also be predicted to have two complementary regions.

Figure 3.19 shows the ^{32}P-labelled 900bp AccI-HindIII restriction fragment (loop DNA) from the subclone 14HH1B hybridised against λmA36 digested with various restriction enzymes. The ^{32}P-labelled probe hybridised to single restriction fragments within digested λmA36, for example the 5.3kb fragment. As discussed above, this fragment includes the region corresponding to the position of the AccI-HindIII fragment of λmA14 and thus λmA36 contains DNA homologous to the probe at equivalent locations in the two clones.

3.1.4 Partial sequencing of λmA14 and λmA36

The mapping of λmA14 and λmA36 has suggested that their similarity extended at least for 11.0kb (with respect to λmA14) from an apparently common KpnI site left of the actin-like region to a BamHI site to the right.

Figure 3.18 Hybridisation of AccI fragment from subclone 14HH1B against digested λmA36

The hybridisation was performed as described in Figure 3.17.

(a) Photograph of a stained gel (lanes 1-6)

(b) Autoradiograph of the nitrocellulose (lanes 1-6)

The fragment(s) which hybridised to the ^{32}P-labelled AccI fragment, from the subclone 14HH1B (Figure 3.16) are indicated below :

Lane	DNA	Restriction enzyme	Hybridised fragment(s) (kb)
1	λmA36	HindIII	2.3 and 5.3
2	λmA36	BamHI	3.0 and 15.7
3	λcI_{857}	HindIII	-
4	pmS4	TaqI	-
5	14HH1B	SstI/AccI	0.9
6	λmA36	BglII	4.3 and 4.7

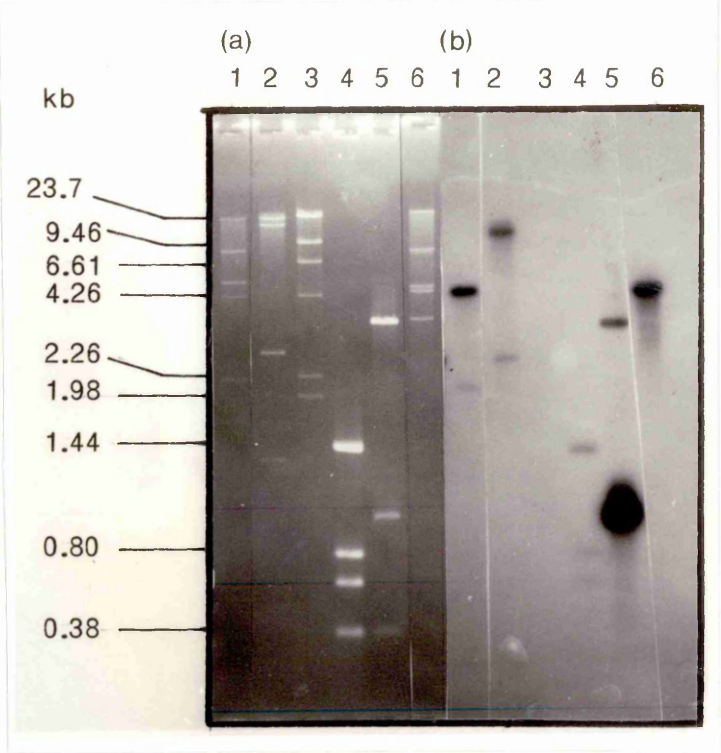

(a)
(b)
1 2 3 4 5 6 1 2 3 4 5 6
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

Figure 3.19 Hybridisation of AccI-HindIII fragment from subclone 14HH1B, against digested λmA36

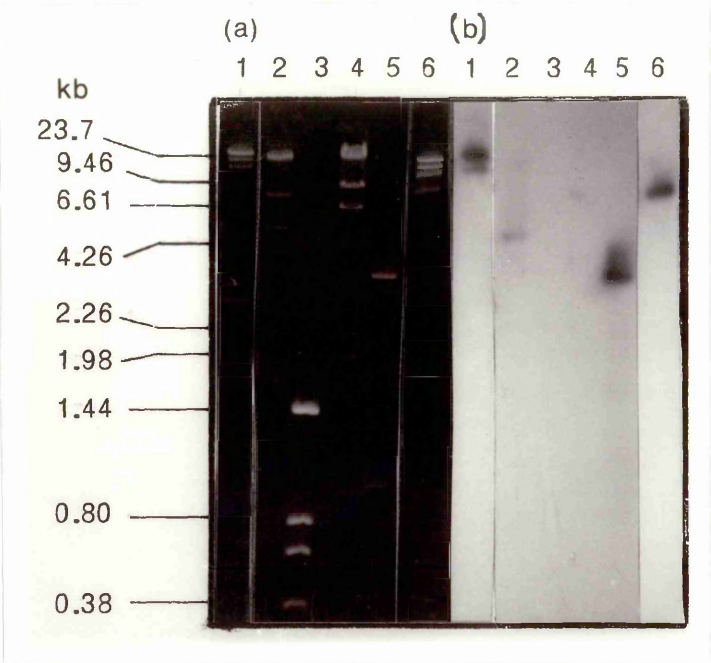
The hybridisation was performed as described in Figure 3.17.

(a) Photograph of the stained gel (lanes 1-6)

(b) Autoradiograph of the nitrocellulose (lanes 1-6)

The fragment(s) which hybridised to the ^{32}P-labelled AccI-HindIII fragment from the subclone 14HH1B (Figure 3.16) are indicated below :

Lane	DNA	Restriction enzyme	Hybridised fragment(s) (kb)
1	λmA36	BamHI	15.7
2	λmA36	HindIII	5.3
3	pmS4	TaqI	-
4	λcI_{857}	HindIII	-
5	14HH1B	SstI/AccI	3.6
6	λmA36	KpnI	12.5


(a)
(b)
1 2 3 4 5 6 1 2 3 4 5 6
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

Because of the insertion in λmA36, and the subsequent displacement sites 5' to the actin pseudogene, it was important to confirm the equivalence of this region 5' to the pseudogene by sequencing from the KpnI site. The clones were also sequenced from the HindIII site 11.0kb to the right to confirm the similarity at the other extremity.

DNA sequence was determined from the extreme 5' KpnI site of the subclones 14KK1 and 36KK1 and from the extreme 3' HindIII site of 14HH4 and 36HH4, as indicated in Figure 3.8. Comparison of sequences from λmA14 and λmA36 is shown in Figure 3.20. It cas be seen that these sequences, although containing some differences are over 96% identical.

Further comparison of λmA14 and λmA36 involved additional sequencing of the KK1 subclones. The subclone 14KK1 was completely sequenced except for 5 bases either side of the AvaII site and 36KK1 was partially sequenced. Figure 3.21 outlines the details of the sequencing of 14KK1 and the DNA sequence obtained is shown in Figure 3.22. The strategy by which 36KK1 was partially sequenced is shown in Figure 3.23. Figure 3.24 shows the DNA sequence obtained. Figure 3.25 shows a comparison of the sequence obtained from 14KK1 and 36KK1 : part (a) is a comparison of the 5' flanking sequences and part (b) is a comparison of the actin-like pseudo-coding sequences.

Figure 3.20 Comparison of the nucleotide sequence of λmA14 and λmA36 at the extremities of corresponding subcloned regions

Sequences shown are from :

(a) The leftward KpnI sites of the subclones 14KK1 and 36KK1 (Figures 3.8 and 3.9).

(b) The rightward HindIII sites of the subclones 14HH4 and 36HH4 (Figures 3.8, 3.11 and 3.12).

```
(a)     'Leftward' extremity (subclones 14KK1 and 36KK1)

          Kpn I
>mA14:  GGTACCAATAGCAGTTAAGGAACGTTCAACATGTCTTAATTTTTCAATAACTTTTCTCCTTATTTTTCTGTTTCAGAGAGTACCTGATTAAAGTATGCCC  100
        ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
>mA36:  GGTACCAATAGCAGTTAAGGAACGTTCAGCATGTCTTAATTTTTCGATAACTTTTCTCCTTATTTTTCTTTTTCAGAGAGTACCTGATTAAAGTATGTCC  100

>mA14:  TTCAAATGATTAGATCAACGAATCAATGTTGATTGTCTATACTATTCCAATAAAATTTTCAGCATGCAATTTCTGAGTGTTGTCTGTGTTTCTTAGTAAG  200
        || ||||||||||||||||||||||||||||||||||||||||| | ||||| ||||||||||||||||||||||||||||||||||||||||||||||||
>mA36:  TTGAAATGATTAGATCAACGAATCAATGTTGATTGTCTGTCCTATTTCAATAAAATTTTCAGCATGCAATTTCTGAGTGTTGTCTGTGTTTCTTAGTAAG  200

>mA14:  GGAGGGGAGAGG                                                                                          212
        ||||||||||||
>mA36:  GGAGGGGAGAGG                                                                                          212

(b)     'Rightward' extremity (subclones 14HH4 and 36HH4)

>mA14:  AGACCGACAGAGACCAACATTGAATAAGGAGGATACTGAGACAATCGGAGGCAATCATGAGGACCTCAAGTTGAGAGAGGACTCGGGGTCACCAGTCTCG  100
        ||||||||||||||| |||||||||||||||| |||||||||||||||||||| |||||||||||| ||||||||||||||||||||||||||||||||
>mA36:  AGACCGACAGAGACCGACATTGAATAAGGAGGACACTGAGACAATCGGAGGCAACCATGAGGACCTCAGGTTGAGAGAGGACTCGGGGTCACCAGTCTCG  100

>mA14:  TAAGAGATGGCCGTTCGAGAGTAGAACGTCCCTTCCACGGGTCTCCAGACCCAGAGGTGAGACAGAGGACGATAACTATTTCAAGGTTTCACTGGGACTG  200
        ||||||| | |||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||
>mA36:  TAAGAGACGTCCGTTCGAGAGTAGAACGTCCCTTCCACGGGCCTCCAGACCCAGAGGTGAGACAGAGGACGATACCTATTTCAAGGTTTCACTGGGACTG  200

>mA14:  GTTCTTCGAA                                                                                           209
        ||||||||||
>mA36:  GTTCTTCGAA                                                                                           209
          Hind III
```

Figure 3.21 Strategy for sequencing the subclone 14KK1

Only those sites used for labelling following primary restriction are shown. The beginning of each arrow denotes the restriction sites at which the fragment was labelled. The arrow tip denotes the limit of the reading of the sequencing gel. The DNA strand designated A corresponds to the sense strand with respect to actin, and strand B represents the antisense strand with respect to actin. The arrows are numbered sequentially along the A and B strands and serve as reference numbers for the table below, outlining the details of the sequencing. The sequence was determined 75% on both strands.

Sequence run	Labelled restriction site	Radionucleotide used	Restriction enzyme second cut	Strand sequenced A or B
1	HindIII*(5'KpnI)	γ^{32}P-ATP	EcoRI	A
2	DraI	γ^{32}P-ATP	BglII	A
3	EcoRI	α^{32}P-dATP	HindIII	A
4	EcoRI	γ^{32}P-ATP	BglII	A
5	HpaII	γ^{32}P-ATP	EcoRI	A
6	AccI	γ^{32}P-ATP	EcoRI	A
7	BglII	α^{32}P-dCTP	HindIII	A
8	BglII	γ^{32}P-ATP	EcoRI	A
9	AvaII	α^{32}P-dCTP	HindIII	A
10	AvaII	γ^{32}P-ATP	EcoRI	A
11	EcoRI*(3'KpnI)	α^{32}P-dATP	HindIII	A
12	HindIII*(5'KpnI)	α^{32}P-dCTP	EcoRI	B
13	DraI	γ^{32}P-ATP	HindIII	B
14	EcoRI	γ^{32}P-ATP	HindIII	B
15	EcoRI	α^{32}P-dATP	BglII	B
16	HpaII	γ^{32}P-ATP	HindIII	B
17	AccI	γ^{32}P-ATP	EcoRI	B
18	BglII	γ^{32}P-ATP	HindIII	B
19	BglII	α^{32}P-dCTP	EcoRI	B
20	AvaII	γ^{32}P-ATP	HindIII	B
21	AvaII	α^{32}P-dCTP	EcoRI	B
22	EcoRI*(3'KpnI)	γ^{32}P-ATP	HindIII	B

* Polylinker restriction site of pUC18

14KK1
Kpn I
Dra I
Eco RI
Hpa II
Acc I
Bgl II
Ava II
Kpn I
100bp
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
A
B

Figure 3.22 Partial nucleotide sequence of 14KK1

The nucleotide sequence of subclone 14KK1 is shown with the leftward KpnI site (Figure 3.8) equivalent to nucleotide 1. Other sites shown are those used in sequencing (Figure 3.21).

```
   Kpn I
   1 GGTACCAATA GCAGTTAAGG AACGTTCAAC ATGTCTTAAT TTTTCAATAA CTTTTCTCCT TATTTTTCTG TTTCAGAGAG TACCTGATTA AAGTATGCCC  100
     CCATGGTTAT CGTCAATTCC TTGCAAGTTG TACAGAATTA AAAAGTTATT GAAAAGAGGA ATAAAAAGAC AAAGTCTCTC ATGGACTAAT TTCATACGGG

 101 TTCAAATGAT TAGATCAACG AATCAATGTT GATTGTCTAT ACTATTCCAA TAAAATTTTC AGCATGCAAT TTCTGAGTGT TGTCTGTGTT TCTTAGTAAG  200
     AAGTTTACTA ATCTAGTTGC TTAGTTACAA CTAACAGATA TGATAAGGTT ATTTTAAAAG TCGTACGTTA AAGACTCACA ACAGACACAA AGAATCATTC

 201 GGAGGGGAGA GGTTTCAGAG TTGGAATGTT CAAGGATACA ACACCTTGGC AAAACACACC AAGAATATGT GCAAATATTT CCGATCTTTT TTCCCCCACA  300
     CCTCCCCTCT CCAAAGTCTC AACCTTACAA GTTCCTATGT TGTGGAACCG TTTTGTGTGG TTCTTATACA CGTTTATAAA GGCTAGAAAA AAGGGGGTGT

 301 ACACGAGATA GAAAGTGAAA ATACTTTATG CCCCTGTAAC TAGAGGATTC TTCCATGAAG TCTGCATTAC AAATCTATGA TATAATATAT AATTTTAGAC  400
     TGTGCTCTAT CTTTCACTTT TATGAAATAC GGGGACATTG ATCTCCTAAG AAGGTACTTC AGACGTAATG TTTAGATACT ATATTATATA TTAAAATCTG
                        Dra I
 401 TCTCTTATTG AATTTTTCTC AAATTTAAAG GAAACTGGGT AGATGTATTG AGGGAATTGA AAACCCGAGT TTTTAACACC GTGATATTCC CCAGTTCATC  500
     AGAGAATAAC TTAAAAAGAG TTTAAATTTC CTTTGACCCA TCTACATAAC TCCCTTAACT TTTGGGCTCA AAAATTGTGG CACTATAAGG GGTCAAGTAG

 501 CGCCAGGTTT GACCTTTCCT TGTCCCATGT TGCATTTTCC GTTCCAATTT TTTTTACCAA AATAAGTGTT CCCACTTTCT TAATATTGCT GAAACGATCT  600
     GCGGTCCAAA CTGGAAAGGA ACAGGGTACA ACGTAAAAGG CAAGGTTAAA AAAAATGGTT TTATTCACAA GGGTGAAAGA ATTATAACGA CTTTGCTAGA
                                                                                                    EcoR I
 601 AGTCAGTAGT CAATTATCCA ACTGCTGTAT AAATGATGAA TGTGTTGTTT TTAACTTGAG CCTATAGATG TGGATGTGGA TAAATTATAG TTGAATTCCA  700
     TCAGTCATCA GTTAATAGGT TGACGACATA TTTACTACTT ACACAACAAA AATTGAACTC GGATATCTAC ACCTACACCT ATTTAATATC AACTTAAGGT

 701 TCTTTTAAAT GCTAGGAATT TCTTCCATGT CTCTCTCTTA CCTGCAATAA ATGCATTTAA AAGAAAGATA AAGTTCTGAC CATTTGTCAA AAAGGATTTC  800
     AGAAAATTTA CGATCCTTAA AGAAGGTACA GAGAGAGAAT GGACGTTATT TACGTAAATT TTCTTTCTAT TTCAAGACTG GTAAACAGTT TTTCCTAAAG

 801 CTACAGCAAG TCATTTGCTG ATGCCATCCT ATGGTATAGG TTGATTTATT TTGCTGATGA TATGGTTTTC TTAAGATTTA TTTATTTTAT ATGAGTACAC  900
     GATGTCGTTC AGTAAACGAC TACGGTAGGA TACCATATCC AACTAAATAA AACGACTACT ATACCAAAAG AATTCTAAAT AAATAAAATA TACTCATGTG

 901 TGTCTAAGTG GTACTAAGAC CCTATTATGG ATGGTGTGTA AGCCAACATG TGGTTGCTTG TGATTGAAAC AAGGACCTCT TGGAAGAGCA GCCAATGATT 1000
     ACAGATTCAC CATGATTCTG GGATAATACC TACCACACAT TCGGTTGTAC ACCAACGAAC ACTAACTTTG TTCCTGGAGA ACCTTCTCGT CGGTTACTAA

1001 TTAACCACTT AGGCATCTCT CCAGCCAGAT TGAAATTATT TTTCATTAGT TGCATTTTTG ATAGGGTCCT ATGGAGACAG GTTAGACTGC AATAGAAGAA 1100
     AATTGGTGAA TCCGTAGAGA GGTCGGTCTA ACTTTAATAA AAAGTAATCA ACGTAAAAAC TATCCCAGGA TACCTCTGTC CAATCTGACG TTATCTTCTT
                                                Hpa II
1101 GAAATCGCCG CACTCGTCAT TGACAATGGC TCCGACCTGC AGGAAGCCGG CTTTGCTGGC GACGACGCCC CCAGGGCCAT GTTCTCTTCC ATCGTAGGGC 1200
     CTTTAGCGGC GTGAGCAGTA ACTGTTACCG AGGCCGTACA CGTTTCGGCC GAAACGACCG CTGCTGCGGG GGTCCCGGTA CAAGAGAAGG TAGCATCCCG
                                                                                  Acc I
1201 GCCCCTGACA CCAGAGTGTC ATGGTGGGCA TGGGCCAGAA AGACTCGTAC GTGGGTGACG AGGCCCAGAG CAAGAGGGGT ATACTGACCC TGAAGTACCC 1300
     CGGGGACTGT GGTCTCACAG TACCACCCGT ACCCGGTCTT TCTGAGCATG CACCCACTGC TCCGGGTCTC GTTCTCCCCA TATGACTGGG ACTTCATGGG
                                      Bgl II
1301 TATCGAACAC GGCATTGTCA CTAACTGGGA CAACATGGAG AAGATCTGGC ACCACACCTT CTACAATGAG CTGCATGTGG CTCCTGAGGA GCCCCGGTAC 1400
     ATAGCTTGTG CCGTAACAGT GATTGACCCT GTTGTACCTC TTCTAGACCG TGGTGTGGAA GATGTTACTC GACGTACACC GAGGACTCCT CGGGGCCATG

1401 TCTGACTGAG GCCCCCTTAA ACCCCAAAGC TAACAGAGAG ATGATGACGC AGATAATATT GGAGATCCTC AATACCCCAG CCATGTACGT GGCCATTCAG 1500
     AGACTGACTC CGGGGGAATT TGGGGTTTCG ATTGTCTCTC TACTACTGCG TCTATTATAA CCTCTAGGAG TTATGGGGTC GGTACATGCA CCGGTAAGTC

1501 GCGGTGCTGT CCTTGTATGC ATCTGGGGAC ACCACTGACA TTGTCATGAA CTCTGGTGAC GGGGTCACAC ACACAGTGCC CATCTAAAAG GGCTACGCCC 1600
     CGCCACGACA GGAACATACG TAGACCCCTG TGGTGACTGT AACAGTACTT GAGACCACTG CCCCAGTGTG TGTGTCACGG GTAGATTTTC CCGATGCGGG
                                      Ava II
1601 TTCCTCACCT CATCTTGCGT CTGGACCTGG CT.....GGA CG.....GAC TGCCTCATGA AGATCCTGAC TAAACGGGGC TACAGCTTTA CCGCCACTGC 1700
     AAGGAGTGGA GTAGAACGCA GACCTGGACC GA.....CCT GG.....CTG ACGGAGTACT TCTAGGACTG ATTTGCCCCG ATGTCGAAAT GGCGGTGACG

1701 TGAGAGGGAA ATTGTTCCTG ACATAAAGGA GAAGCTGTGC TATGTTGCCC TGGATTTTGA GCAAGAAATG GCTACTGCTG CATCATCTTC CTCCTTGGAG 1800
     ACTCTCCCTT TAACAAGGAC TGTATTTCCT CTTCGACACG ATACAACGGG ACCTAAAACT CGTTCTTTAC CGATGACGAC GTAGTAGAAG GAGGAACCTC

1801 AAGAGTTACC AGTTGCCCGA CGGGCAGGCG ATCACCATTG GCAACGAGCG GTTCCGGTGT CCGGAGGCAC TCTTCCAGCC TTCCTTCCTA GGCATAGAGT 1900
     TTCTCAATGG TCAACGGGCT GCCCGTCCGC TAGTGGTAAC CGTTGCTCGC CAAGGCCACA GGCCTCCGTG AGAAGGTCGG AAGGAAGGAT CCGTATCTCA
                                                                                                    Kpn I
1901 CCTGTGGTAT CCATGAGACC ACCTTCAACT CCATCATGAA GTGTGATGTG GATATCCGCA AAGACCTGTA TGCCAAAACA GTGCTGTCTA GCGGTACC  1998
     GGACACCATA GGTACTCTGG TGGAAGTTGA GGTAGTACTT CACACTACAC CTATAGGCGT TTCTGGACAT ACGGTTTTGT CACGACAGAT CGCCATGG
```

Figure 3.23 Strategy for sequencing subclone 36KK1

Only those sites used for labelling following primary restriction are shown. This subclone was partially sequenced from four restriction sites the 5'KpnI, EcoRI, BglII and 3'KpnI and the sequence data frome each of these has been designated (i), (ii), (iii) and (iv) respectively. Further details of this figure are as described for Figure 3.21. The arrows are numbered to serve as a reference for the table below, outlining the details of the sequencing.

Sequence run	Labelled restriction site	Radionucleotide used	Restriction enzyme second cut	Strand sequenced A or B
1	HindIII* (5'KpnI)	γ^{32}P-ATP	EcoRI	A
2	EcoRI	α^{32}P-dATP	HindIII	A
3	EcoRI	γ^{32}P-ATP	BglII	A
4	BglII	α^{32}P-dCTP	HindIII	A
5	BglII	γ^{32}P-ATP	EcoRI	A
6	EcoRI* (3'KpnI)	α^{32}P-dATP	HindIII	A
7	HindIII* (5'KpnI)	α^{32}P-dCTP	EcoRI	B
8	EcoRI	γ^{32}P-ATP	HindIII	B
9	EcoRI	α^{32}P-dATP	BglII	B
10	BglII	γ^{32}P-ATP	HindIII	B
11	BglII	α^{32}P-dCTP	EcoRI	B

* Polylinker restriction site of pUC18

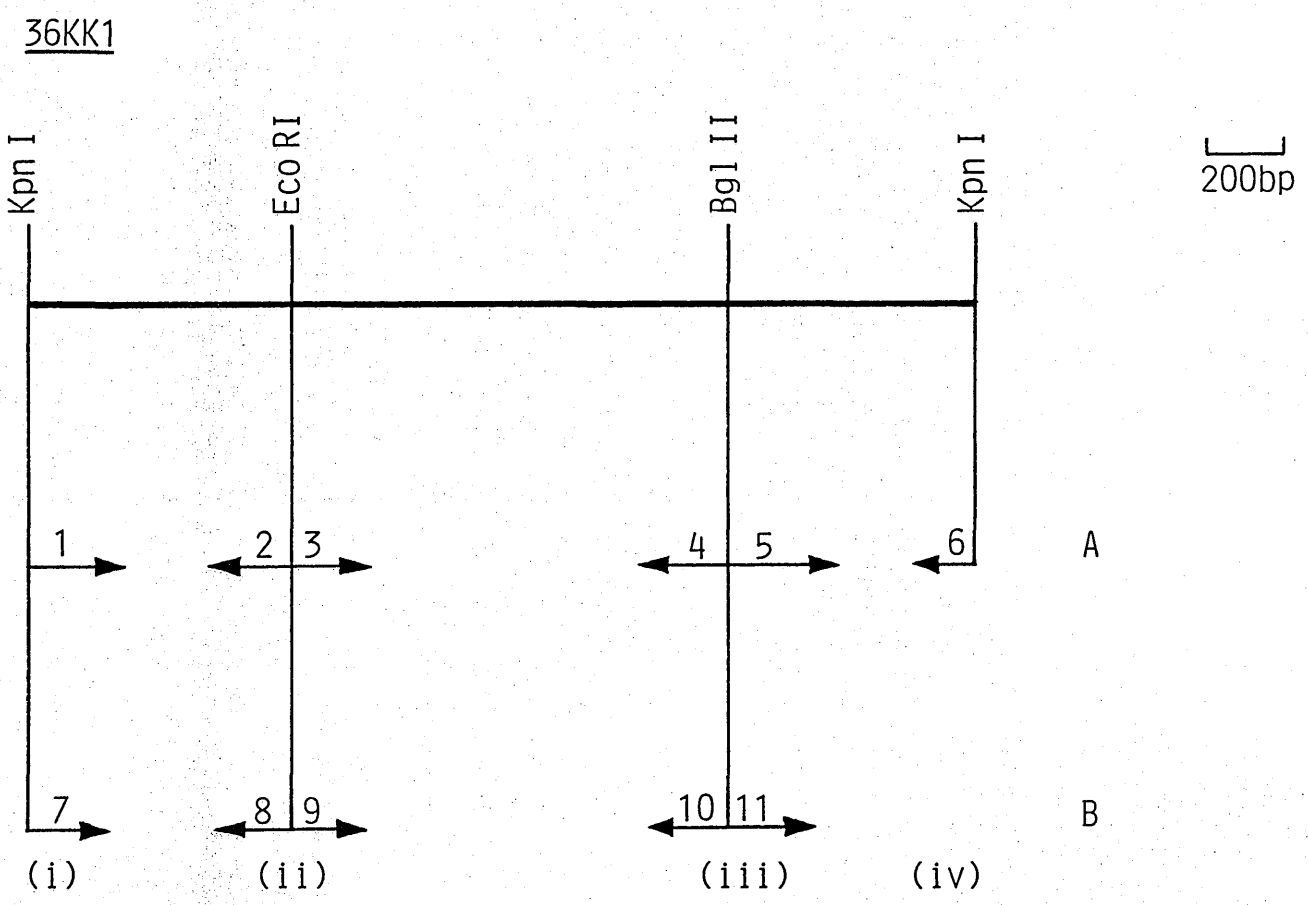

36KK1
Kpn I
Eco RI
Bgl II
Kpn I
200bp
1
2
3
4
5
6
A
7
8
9
10
11
B
(i)
(ii)
(iii)
(iv)

Figure 3.24 Partial nucleotide sequence of subclone 36KK1

The nucleotide sequence of subclone 36KK1 is shown with the leftward KpnI site (Figure 3.8) equivalent to nucleotide 1. The sites shown are those used in sequencing, 5'kpnI, EcoRI, BglII and 3'KpnI and the sequence data from each of these has been designated (i), (ii), (iii) and (iv) respectively (Figure 3.23).

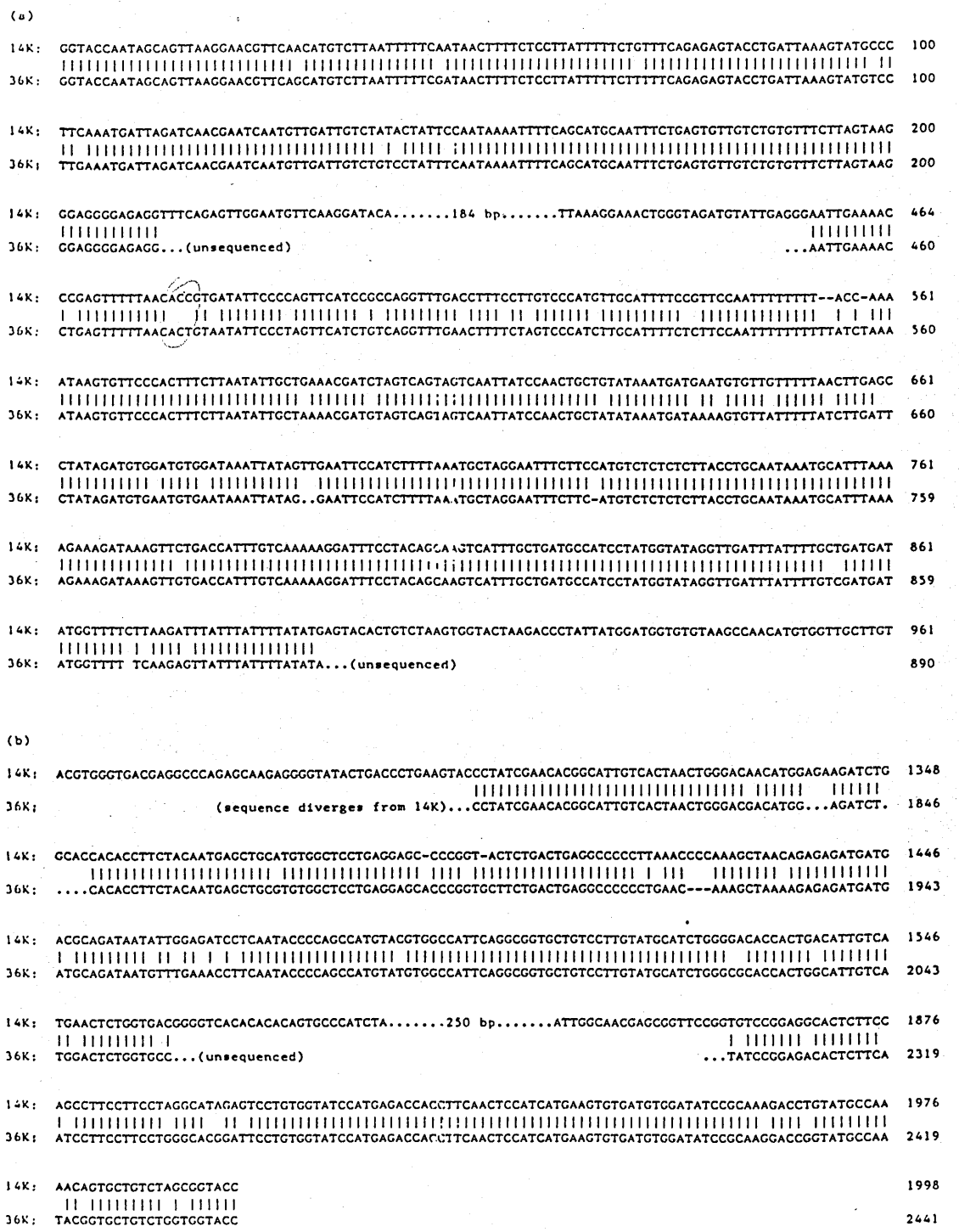
```
(i)
    Kpn I
   1 GGTACCAATA GCAGTTAAGG AACGTTCAGC ATGTCTTAAT TTTTCGATAA CTTTTCTCCT TATTTTTCTT TTTCAGAGAG TACCTGATTA AAGTATGTCC 100
     CCATGGTTAT CGTCAATTCC TTGCAAGTCG TACAGAATTA AAAAGCTATT GAAAAGAGGA ATAAAAAGAA AAAGTCTCTC ATGGACTAAT TTCATACAGG

 101 TTGAAATGAT TAGATCAACG AATCAATGTT GATTGTCTGT CCTATTTCAA TAAAATTTTC AGCATGCAAT TTCTGAGTGT TGTCTGTGTT TCTTAGTAAG 200
     AACTTTACTA ATCTAGTTGC TTAGTTACAA CTAACAGACA GGATAAAGTT ATTTTAAAAG TCGTACGTTA AAGACTCACA ACAGACACAA AGAATCATTC

 201 GGAGGGGAGA GG                                                                                              212
     CCTCCCCTCT CC


(ii)

 451 AATTGAAAAC CTGAGTTTTT AACACTGTAA TATTCCCTAG TTCATCTGTC AGGTTTGAAC TTTTCTAGTC CCATCTTGCA TTTTCTCTTC CAATTTTTTT 550
     TTAACTTTTG GACTCAAAAA TTGTGACATT ATAAGGGATC AAGTAGACAG TCCAAACTTG AAAAGATCAG GGTAGAACGT AAAAGAGAAG GTTAAAAAAA

 551 TTTATCTAAA ATAAGTGTTC CCACTTTCTT AATATTGCTA AAACGATGTA GTCAGTAGTC AATTATCCAA CTGCTATATA AATGATAAAA GTGTTATTTT 650
     AAATAGATTT TATTCACAAG GGTGAAAGAA TTATAACGAT TTTGCTACAT CAGTCATCAG TTAATAGGTT GACGATATAT TTACTATTTT CACAATAAAA
                                                 EcoR I
 651 TATCTTGATT CTATAGATGT GAATGTGAAT AAATTATAG. .GAATTCCAT CTTTTAAATG CTAGGAATTT CTTCATGTCT CTCTCTTACC TGCAATAAAT 750
     ATAGAACTAA GATATCTACA CTTACACTTA TTTAATATC. .CTTAAGGTA GAAAATTTAC GATCCTTAAA GAAGTACAGA GAGAGAATGG ACGTTATTTA

 751 GCATTTAAAA GAAAGATAAA GTTGTGACCA TTTGTCAAAA AGGATTTCCT ACAGCAAGTC ATTTGCTGAT GCCATCCTAT GGTATAGGTT GATTTATTTT 850
     CGTAAATTTT CTTTCTATTT CAACACTGGT AAACAGTTTT TCCTAAAGGA TGTCGTTCAG TAAACGACTA CGGTAGGATA CCATATCCAA CTAAATAAAA

 851 GTCGATGATA TGGTTTTTCA AGAGTTATTT ATTTTATATA                                                                 890
     CAGCTACTAT ACCAAAAAGT TCTCAATAAA TAAAATATAT


(iii)

1671 CGGGGAAGGC AGAGCACAGG GAGTGAAGAA CTACCCTTGG CACATGCGCA GATTATTTGT TTACCAATTA GAACACAGGA TGTCAGCACC ATCTTGCAAC 1770
     GCCCCTTCCG TCTCGTGTCC CTCACTTCTT GATGGGAACC GTGTACGCGT CTAATAAACA AATGGTTAAT CTTGTGTCCT ACAGTCGTGG TAGAACGTTG
                                                                                  Bgl II
1771 GGTGAATGTG AGGGCGGCTT CCCACACCTA TCGAACACGG CATTGTCACT AACTGGGACG ACATGG...A GATCT..... CACACCTTCT ACAATGAGCT 1870
     CCACTTACAC TCCCGCCGAA GGGTGTGGAT AGCTTGTGCC GTAACAGTGA TTGACCCTGC TGTACC...T CTACA..... GTGTGGAAGA TGTTACTCGA

1871 GCGTGTGGCT CCTGAGGAGC ACCCGGTGCT TCTGACTGAG GCCCCCCTGA ACAAAGCTAA AAGAGAGATG ATGATGCAGA TAATGTTTGA AACCTTCAAT 1970
     CGCACACCGA GGACTCCTCG TGGGCCACGA AGACTGACTC CGGGGGGACT TGTTTCGATT TTCTCTCTAC TACTACGTCT ATTACAAACT TTGGAAGTTA

1971 ACCCCAGCCA TGTATGTGGC CATTCAGGCG GTGCTGTCCT TGTATGCATC TGGGCGCACC ACTGGCATTG TCATGGACTC TGGTGCC               2057
     TGGGGTCGGT ACATACACCG GTAAGTCCGC CACGACAGGA ACATACGTAG ACCCGCGTGG TGACCGTAAC AGTACCTGAG ACCACGG


(iv)

2301 TATCCGGAGA CACTCTTCAA TCCTTCCTTC CTGGGCACGG ATTCCTGTGG TATCCATGAG ACCACCTTCA ACTCCATCAT GAAGTGTGAT GTGGATATCC 2400
     ATAGGCCTCT GTGAGAAGTT AGGAAGGAAG GACCCGTGCC TAAGGACACC ATAGGTACTC TGGTGGAAGT TGAGGTAGTA CTTCACACTA CACCTATAGG
                                    Kpn I
2401 GCAAGGACCG GTATGCCAAT ACGGTGCTGT CTGGTGGTAC C                                                             2441
     CGTTCCTGGC CATACGGTTA TGCCACGACA GACCACCATG G
```

Figure 3.25 Comparison of the nucleotide sequence from subclones 14KK1 and 36KK1

The regions of 14KK1 and 36KK1 compared are :

(a) Regions 5' to the actin pseudogene

(b) Regions within the actin pseudogene coding sequence



3.2 Analysis of the foldback structure in λmA14

The second part of this work was centred on the foldback structure in λmA14, the main thrust being nucleotide sequence determination. As it had been established above, that λmA14 and λmA36 had similar overall structures over much of the foldback region, it was decided to concentrate on a single clone, and the larger and more complete clone, λmA14, was chosen.

3.2.1 Location of the inverted repeat DNA of the stem

Before structural analysis could be undertaken it was necessary to locate within the restriction map of λmA14 the inverted repeat DNA of the stem visualised by the electron microscopy. From electron micrographic measurements the left-hand arm of the foldback structure in λmA14 was estimated to occur within 50 nucleotides of the 3' non-coding actin-like region and would therefore be expected to be contained within the subclone 14HH1, as shown in Figure 3.26. The subclone 14HH1A was partially sequenced to the left of the SstI site as indicated in Figure 3.32, gel runs 1 and 13. The SstI site was predicted on the basis of the electron micrographs and restriction mapping to occur beyond the end of the 3' non-coding actin pseudogene and possibly within the left-hand arm DNA of the foldback. In Figure 3.27 the sequence from the left of the SstI site was compared with the 3' non-coding end of the actin processed pseudogene in λmA19 (Leader *et al.*,1985). It can be seen that the homology to the γ-actin pseudogene begins 130bp to the left of the SstI site. Therefore on the basis of the measurements mentioned above it was assumed that the SstI site fell just within the DNA of

Figure 3.26 Location of DNA probes derived from subclone 14HH1B used to analyse the foldback structure within λmA14

The figure indicates the origin of the three DNA probes (i), (ii) and (iii) used to hybridise to digested λmA14 and the DNA probe (iv) used to hybridise to digested mouse DNA :

(i) SstI-AccI 400bp fragment

(ii) AccI 1.0kb fragment

(iii) AccI-HindIII 900bp fragment

(iv) SstI-AvaII 820bp fragment

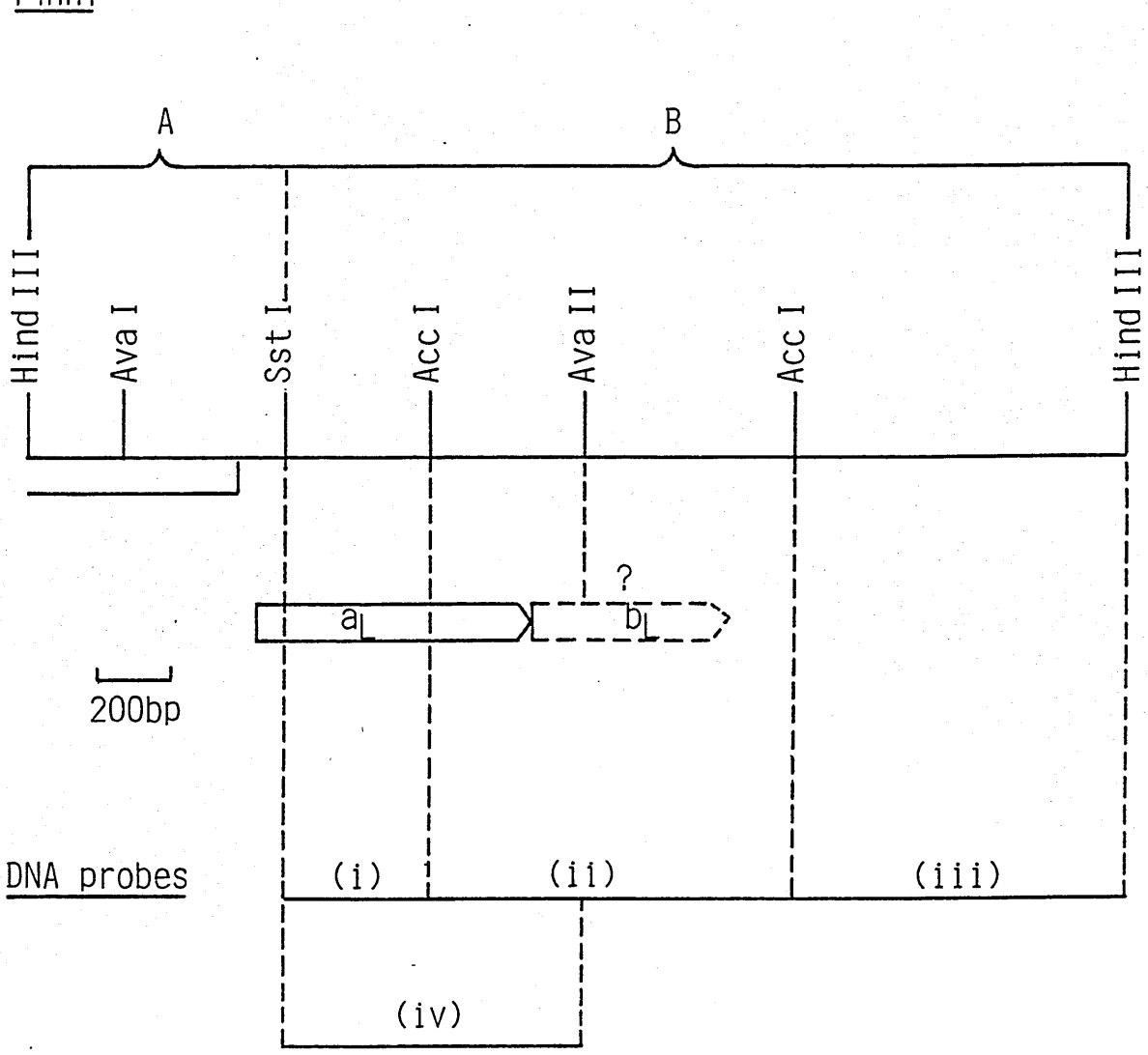
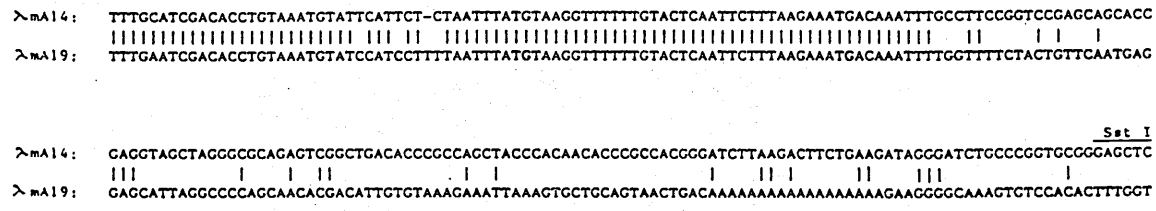


Figure 3.27 Comparison of the 14HH1A nucleotide sequence with the 3'end of the actin pseudogene in λmA19

The nucleotide sequence from the SstI site within 14HH1A (Figure 3.32, sequence runs 1 and 13) is compared with the nucleotide sequence from the 3' non-coding end of the actin processed pseudogene in λmA19 (Leader *et al.*, 1985).



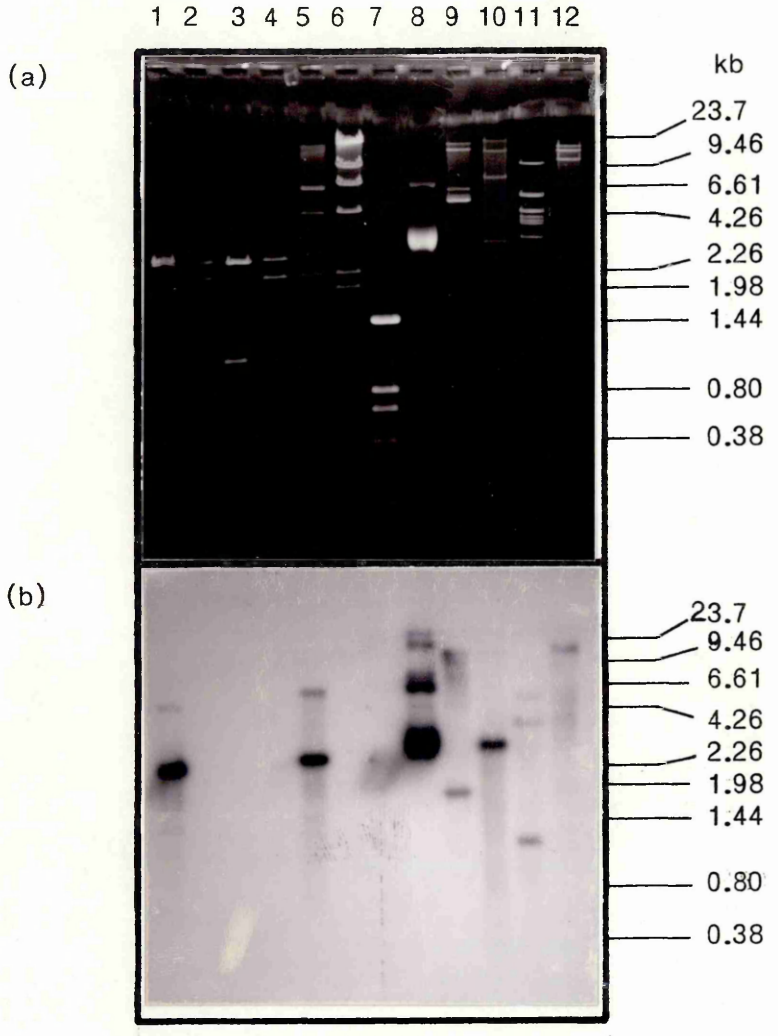
the left-hand arm of the foldback. It should be pointed out that the end of the actin-like sequence described was 100bp from the presumed 3' end of the corresponding actin mRNA, as judged by comparison with the position of the 3' poly A tail of λmA19. The question of the apparent truncation of the actin pseudogene in λmA14, is addressed in section 3.2.6.

The DNA to the right of the SstI site contained within the subclone 14HH1B was therefore used to locate the DNA of the right-hand arm of the foldback by hybridising against digested λmA14. Part of the a_L region (760bp) of λmA14 was predicted to occur within a 400bp SstI-AccI fragment within the subclone 14HH1B, shown in Figure 3.26. This restriction fragment was used to locate the DNA complementary to a_L within λmA14, that is the right-hand arm of the foldback designated a_R, in Figure 1.8. Figure 3.28 shows the ^{32}P-labelled SstI-AccI restriction fragment hybridised against the subclones of λmA14 (14HH1, 14HH2, 14HH3 and 14HH4) digested with HindIII, and λmA14 digested with various restriction enzymes. The SstI-AccI probe did not hybridise to the λmA14 HindIII subclones, 14HH2, 14HH3 and 14HH4. However the SstI-AccI fragment did hybridise to restriction fragments which occurred beyond the region of the λmA14 HindIII subclones, for example, a 6.5kb HindIII and 1.9kb SmaI restriction fragment. The 1.9kb SmaI restriction fragment occurred more than 11.0kb to the right of the location of the SstI-AccI fragment (a_L region). The electron micrograph measurements had predicted the right-hand arm of the foldback complementary to a_L (designated a_R in Figure 1.8) would occur approximately 11.0kb from a_L, and therefore was consistent with the assignment of a_L to this SmaI fragment

Figure 3.28 Analysis of λmA14 foldback structure by hybridisation of SstI-AccI fragment from subclone 14HH1B, against digested λmA14

The ^{32}P-labelled SstI-AccI restriction fragment from the subclone 14HH1B (Figure 3.26) was hybridised against λmA14 HindIII subclones digested with HindIII and λmA14 digested with various restriction endonucleases. The length of the fragment(s) which hybridise to the DNA probe are indicated below :

Lane	DNA	Restriction enzyme	fragment(s) kb
1	14HH1	HindIII	3.0
2	14HH2	HindIII	-
3	14HH3	HindIII	-
4	14HH4	HindIII	-
5	λmA14	HindIII	3.0 and 6.5
6	λcI_{857}	HindIII	-
7	pmS4	TaqI	-
8	14HH1B	-	-
9	λmA14	SmaI	1.9 and 16.5
10	λmA14	SstI	3.2
11	λmA14	PvuII	1.2, 3.0 and 4.5
12	λmA14	KpnI	15.3


1 2 3 4 5 6 7 8 9 10 11 12
(a)
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38
(b)
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

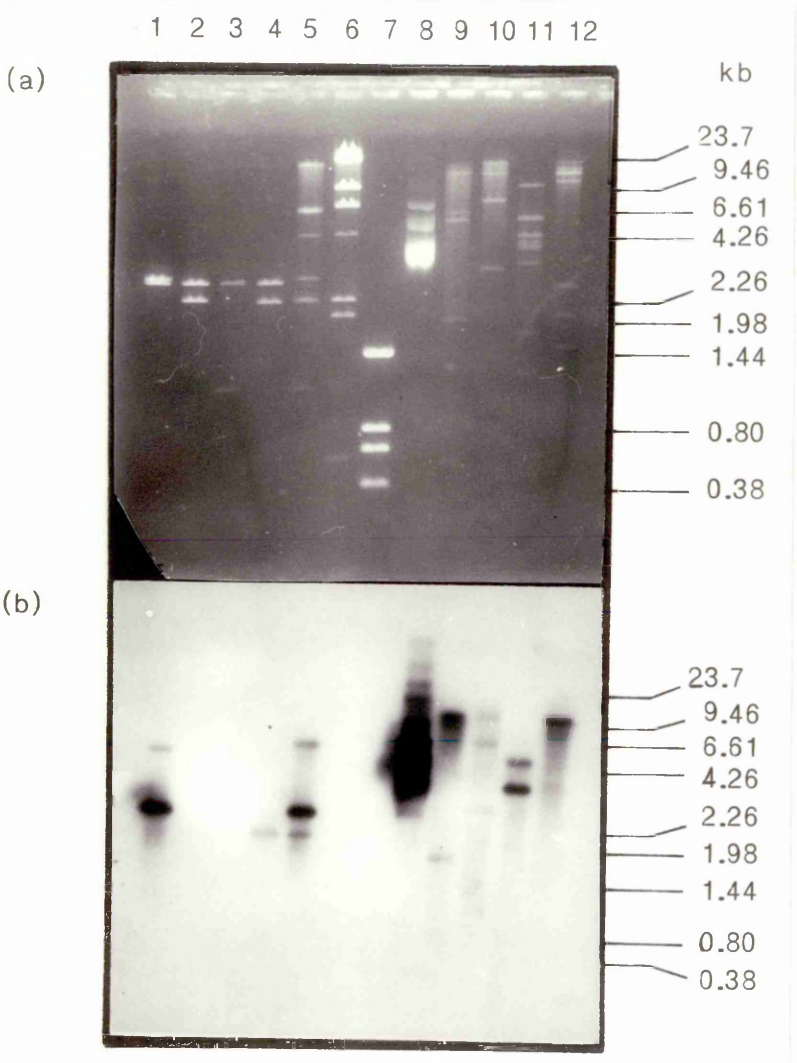
which was therefore subcloned as 14SS1 (Figure 3.14).

Hybridisation was then performed using a 1.0kb AccI restriction fragment from 14HH1B, shown in Figure 3.26. If the upper left-hand arm of the foldback, designated b_L, in Figure 1.8, were located directly adjacent to a_L as shown in Figure 1.8 (a), this restriction fragment would also contrain 540bp of b_L DNA and be expected to hybridise to a region corresponding to b_R in addition to hybridising to a_R. If the AccI fragment only contained the a_L region it would only hybridise to a_R. Figure 3.29 shows the ^{32}P-labelled AccI restriction fragment was hybridised against λmA14 HindIII subclones digested with HindIII and λmA14 digested with various restriction enzymes. The AccI restriction fragment hybridised to the subclone 14HH4 (2.3kb HindIII fragment) which occurred 5.2kb to the right. The AccI fragment also hybridised to the same restriction fragments observed for the SstI-AccI restriction fragment, those which occurred beyond the 14HH4. This confirmed that the AccI restriction fragment contained part of the a_L and b_L regions. The SstI-AccI fragment which had contained only part of the a_L region had not hybridised to the subclone 14HH4 (2.3kb HindIII fragment). Therefore the 2.3kb HindIII restriction fragment must contain the DNA complementary to b_L, (designated b_R, Figure 1.8). In one of the two alternative interpretations of the electron micrograph measurements, Figure 1.8(a), b_R was positioned 5.2kb to the right of b_L, consistent with it being contained in the 2.3kb HindIII fragment that had been subcloned into 14HH4 (Figure 3.13). The alternative arrangement shown in Figure 1.8(b) would have had both b_L and b_R at different positions. The arrangement of λmA14

Figure 3.29 Analysis of λmA14 foldback structure by hybridisation of AccI fragment from subclone 14HH1B, against digested λmA14

The ^{32}P-labelled AccI restriction fragment from the subclone 14HH1B (Figure 3.26) was hybridised against λmA14 HindIII subclones digested with HindIII and λmA14 digested with various restriction endonucleases. The length of the fragment(s) which hybridise to the DNA probe are indicated below :

Lane	DNA	Restriction enzyme	Hybridised fragment(s) (kb)
1	14HH1	HindIII	3.0
2	14HH2	HindIII	-
3	14HH3	HindIII	-
4	14HH4	HindIII	2.3
5	λmA14	HindIII	2.3, 3.0 and 6.5
6	λcI_{857}	HindIII	-
7	pmS4	TaqI	-
8	14HH1B	-	-
9	λmA14	SmaI	1.9 and 16.5
10	λmA14	SstI	3.2, 7.5 and 14.5
11	λmA14	PvuII	-
12	λmA14	KpnI	15.3


1 2 3 4 5 6 7 8 9 10 11 12
(a)
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38
(b)
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

shown in Figure 1.8(a) was therefore concluded to be correct.

The 900bp AccI-HindIII restriction fragment from 14HH1 (shown in Figure 3.26) was predicted from the electron micrograph measurements to contain the DNA of the main loop in λmA14 and no repetitive stem DNA. Figure 3.30 shows the ^{32}P-labelled AccI-HindIII restriction fragment hybridised against λmA14 HindIII subclones digested with HindIII, and λmA14 digested with various restirction enzyme. The AccI-HindIII restriction fragment only hybridised to subclone 14HH1, (from which it was derived) and to single restriction fragment of digested λmA14. The results confirmed that b_L was totally contained in the AccI fragment of 14HH1, and did not extend into the AccI-HindIII fragment.

Figure 3.31 summaries the relationship of the electron micrograph stem sections to the restriction map of λmA14, the subclones containing these stem section being indicated.

3.2.2 Sequencing the subclones containing the stem DNA

The subclones containing the stem DNA of the foldback structure in λmA14 were sequenced in whole or in part.

Figure 3.32 outlines the details of the partial sequencing of the subclone 14HH1, which contains the left-hand arm of the stem (a_L and b_L regions). This sequence was designated LH and is shown in Figure 3.33.

Figure 3.34 outlines the details of the sequencing of the subclone 14SS1, which contains the lower right-hand arm of the stem (a_R). This sequence was designated RH1 and is shown in Figure 3.35.

Figure 3.36 outlines the details for the partial sequencing of the

Figure 3.30 Analysis of λmA14 foldback structure by hybridisation of AccI-HindIII fragment from subclone 14HH1B, against digested λmA14

The ^{32}P-labelled AccI-HindIII restriction fragment from the subclone 14HH1B (Figure 3.26) was hybridised against λmA14 HindIII subclones digested with HindIII and λmA14 digested with various restriction endonucleases. The length of the fragment(s) which hybridise to the DNA probe are indicated below :

Lane	DNA	Restriction enzyme	Hybridised fragment(s) (kb)
1	14HH1	HindIII	3.0
2	14HH2	HindIII	-
3	14HH3	HindIII	-
4	14HH4	HindIII	-
5	λmA14	HindIII	3.0
6	λcI_{857}	HindIII	-
7	pmS4	TaqI	-
8	14HH1B	-	-
9	λmA14	SmaI	16.5
10	λmA14	SstI	3.2
11	λmA14	PvuII	3.5
12	λmA14	KpnI	15.3

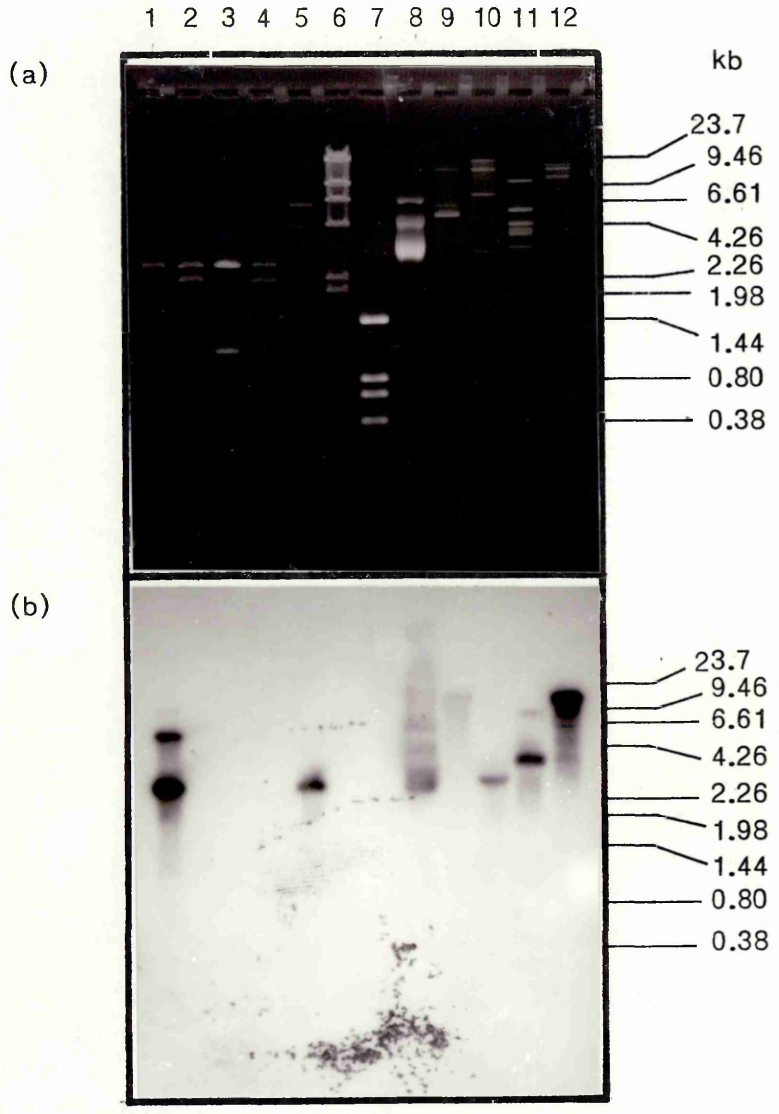

1 2 3 4 5 6 7 8 9 10 11 12
(a)
kb
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38
(b)
23.7
9.46
6.61
4.26
2.26
1.98
1.44
0.80
0.38

Figure 3.31 Relationship of electron microscopic stem sections to λmA14

The diagram shows the various stem sections identified from the electron micrographs, positioned in the various subclones of λmA14 on the basis of the hybridisation results of Figures 3.28 to 3.30.

λmA14

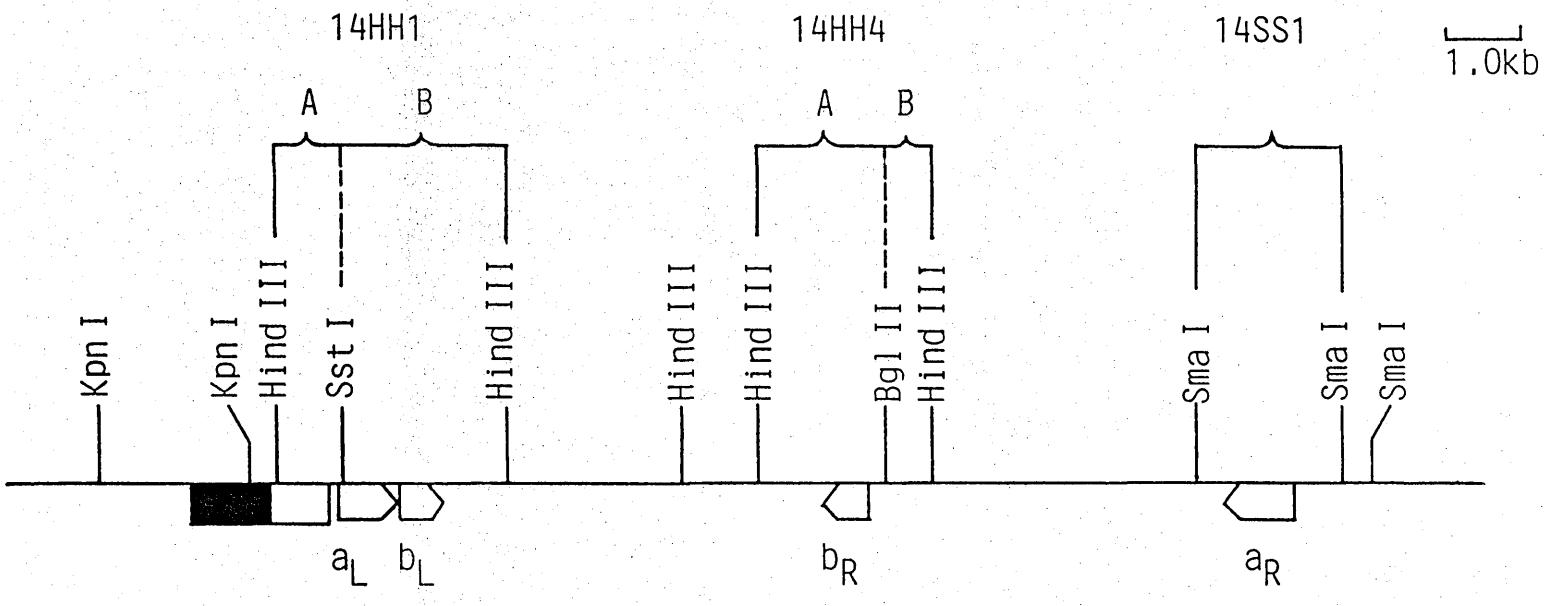

14HH1
14HH4
14SS1
1.0kb
A
B
A
B
Kpn I
Kpn I
Hind III
Sst I
Hind III
Hind III
Hind III
Bgl II
Hind III
Sma I
Sma I
Sma I
a_L
b_L
b_R
a_R

Figure 3.32 Strategy for sequencing subclone 14HH1

Only those sites used for labelling following primary restriction are shown. The details of this figure are as described for Figure 3.21. The arrows are numbered to serve as a reference for the table below, outlining the details of the sequencing.

Sequence run	Labelled restriction site	Radionucleotide used	Restriction enzyme second cut	Strand sequenced A or B
1	EcoRI* (SstI)	α^{32}P-dATP	HindIII	A
2	EcoRI* (SstI)	γ^{32}P-ATP	HindIII	A
3	HpaII (a)	γ^{32}P-ATP	AvaII	A
4	AccI (a)	γ^{32}P-ATP	HpaII	A
5	AvaII (a)	α^{32}P-dCTP	EcoRI	A
6	HpaII (b)	α^{32}P-dCTP	AccI	A
7	HpaII (b)	γ^{32}P-ATP	HindIII	A
8	AccI (b)	γ^{32}P-ATP	HindIII	A
9	TaqI	α^{32}P-dCTP	EcoRI	A
10	TaqI	γ^{32}P-ATP	HindIII	A
11	AvaII (b)	γ^{32}P-ATP	HindIII	A
12	HindIII	α^{32}P-dCTP	EcoRI	A
13	EcoRI* (SstI)	γ^{32}P-ATP	HindIII	B
14	EcoRI* (SstI)	α^{32}P-dATP	HindIII	B
15	HpaII (a)	γ^{32}P-ATP	EcoRI	B
16	AccI (a)	γ^{32}P-ATP	EcoRI	B
17	AvaII (a)	γ^{32}P-ATP	EcoRI	B
18	HpaII (b)	γ^{32}P-ATP	AccI	B
19	HpaII (b)	α^{32}P-dCTP	HindIII	B
20	AccI (b)	γ^{32}P-ATP	AvaII	B
21	TaqI	γ^{32}P-ATP	EcoRI	B
22	TaqI	α^{32}P-dCTP	HindIII	B
23	AvaII (b)	γ^{32}P-ATP	AccI	B

* Polylinker restriction site of pUC18

14HH1

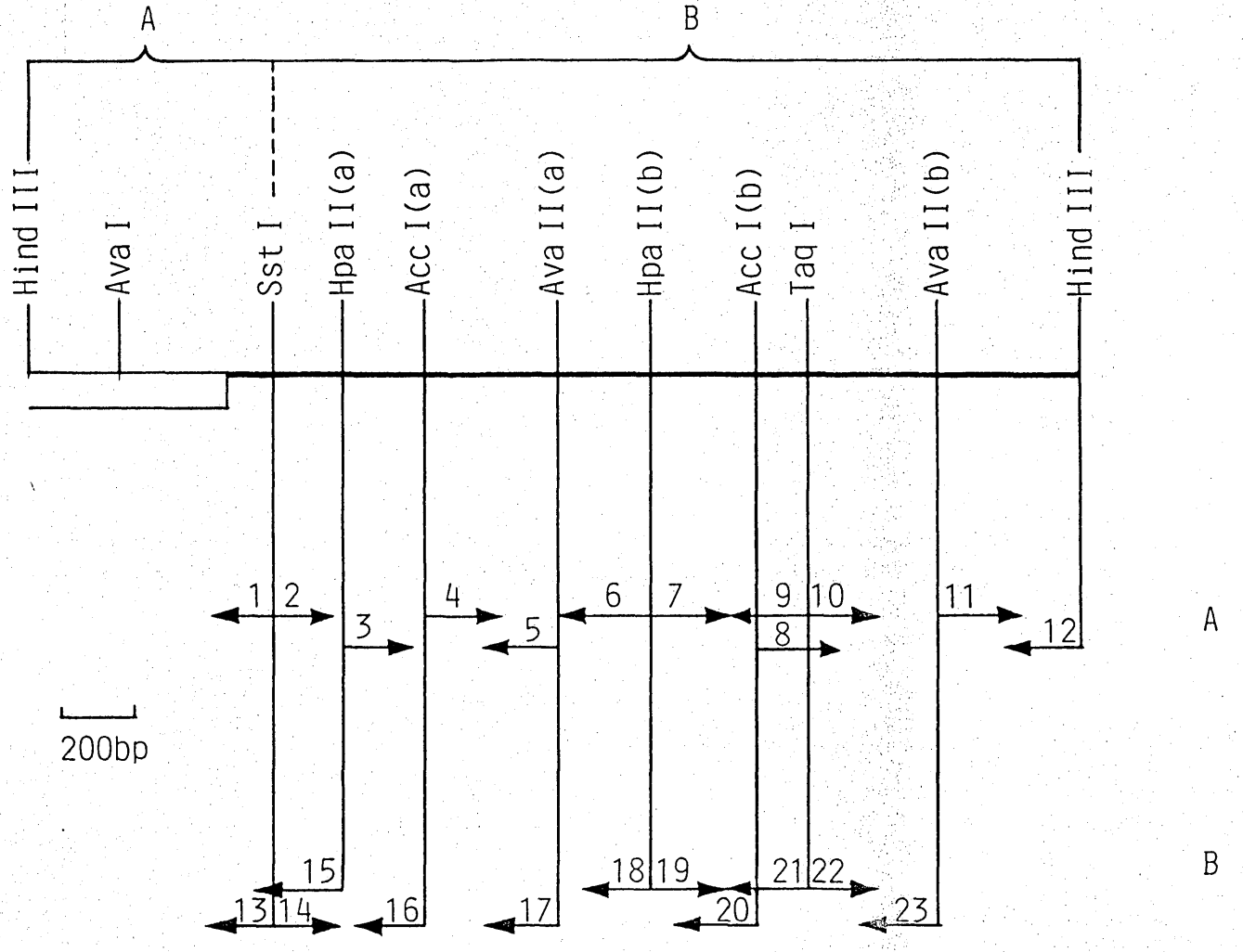

A
B
Hind III
Ava I
Sst I
Hpa II (a)
Acc I (a)
Ava II (a)
Hpa II (b)
Acc I (b)
Taq I
Ava II (b)
Hind III
1
2
3
4
5
6
7
8
9
10
11
12
200bp
13
14
15
16
17
18
19
20
21
22
23
A
B

Figure 3.33 Partial nucleotide sequence of subclone 14HH1

The nucleotide sequence is numbered from the first base that diverged from the 3' non-coding actin DNA, and is designated LH in the text as it contains the left-hand portion of the foldback stem (Figure 3.31).

```
                Ava II
   1 GCCTTCCGGT CCGAGCAGCA CCGAGGTAGC TAGGGCGCAG AGTCGGCTGA CACCCGCCAG CTACCCACAA CACCCGCCAC GGGATCTTAA GACTTCTGAA  100
     CGGAAGGCCA GGCTCGTCGT GGCTCCATCG ATCCCGCGTC TCAGCCGACT GTGGGCGGTC GATGGGTGTT GTGGGCGGTG CCCTAGAATT CTGAAGACTT

                          Sst I
 101 GATAGGGATC TGCCCCGTGC GGGAGCTCTT TGCCTGAGAA TCAGCAGCAG ACATCTTGGT TCCAGGACTC CACCGAGTGT ATCCTGCACA GCTCCAGAGA  200
     CTATCCCTAG ACGGGGCACG CCCTCGAGAA ACGGACTCTT AGTCGTCGTC TGTAGAACCA AGGTCCTGAG GTGGCTCACA TAGGACGTGT CGAGGTCTCT
        Pvu II
 201 ATACCAGCTG GCTAAAGGCA AACGTAAGAA TCCTACTAAC AGAAATCAAG ACCAATCACC ATCACCAGGA CGCAGCACTC CCAACCCCAC CTAGTCCTGT  300
     TATGGTCGAC CGATTTCCGT TTGCATTCTT AGGATGATTG TCTTTAGTTC TGGTTAGTGG TAGTGGTCCT GCGTCGTGAG GGTTGGGGTG GATCAGGACA
                                Hpa II
 301 GCACCCCAAC ACAACCGAAA AGTCAAGAAC CCGGAATTAA AGCATATCTC ATTATGATGG TAGAGGACAT CAAGAAGGAC TTTAATAACT CACTTAAAGA  400
     CGTGGGGTTG TGTTGGCTTT TCAGTTCTTG GGCCTTAATT TCGTATAGAG TAATACTACC ATCTCCTGTA GTTCTTCCTG AAATTATTGA GTGAATTTCT

 401 AATACAGGAG AACACTGCTA ACGAGTTACA AGTTCTTAAA GAAAAACAGG AAAACACAAC CAAACAGGTA GAAGTCCTTA AAGAAAAACA GGAAAACACA  500
     TTATGTCCTC TTGTGACGAT TGCTCAATGT TCAAGAATTT CTTTTTGTCC TTTTGTGTTG GTTTGTCCAT CTTCAGGAAT TTCTTTTTGT CCTTTTGTGT
                                                          Acc I
 501 TCCAAACAGG TGATGGAAAT GAACAAAACC ATACTAGACC TAAAAAGGGA AGTAGACATC AATAAAGAAA ACCCAAAGTG AGGCAACGCT GGAGTTAGAA  600
     AGGTTTGTCC ACTACCTTTA CTTGTTTTGG TATGATCTGG ATTTTTCCCT TCATCTGTAG TTATTTCTTT TGGGTTTCAC TCCGTTGCGA CCTCAATCTT

 601 ACCCTAGGAA AGAAATCTGG AACCATAGAT GCGAGCATCA GGAACAGAAT ACAAGAGATG GAAGAGAGAA TCTCAGGTGC AGAAGATTCC ATAGAGAACA  700
     TGGGATCCTT TCTTTAGACC TTGGTATCTA CGCTCGTAGT CCTTGTCTTA TGTTCTCTAC CTTCTCTCTT AGAGTCCACG TCTTCTAAGG TATCTCTTGT

 701 TCGGCACAAC AATCAAAGAA AATACAAAAT GCAGAAGGAG CCTAACTCAA AACATTCAGG AAATACAGGA CACAATGAGA AGACCAAACC TACAGATAAC  800
     AGCCGTGTTG TTAGTTTCTT TTATGTTTTA CGTCTTCCTC GGATTGAGTT TTGTAAGTCC TTTATGTCCT GTGTTACTCT TCTGGTTTGG ATGTCTATTG

 801 AGGAGTTGAT GAGAATGAAG ATTTTCAACT TAAAGGGCCA GCAAATATAT TCAACAAAAT TATAGAAGAA AACTTCCCAA ACCTAAAGAA AGAAATGCCC  900
     TCCTCAACTA CTCTTACTTC TAAAAGTTGA ATTTCCCGGT CGTTTATATA AGTTGTTTTA ATATCTTCTT TTGAAGGGTT TGGATTTCTT TCTTTACGGG
                                Ava II
 901 ATGAATATAC AGGAAGCCTA CAGAACTCCA AATAGACTGG ACCAGAAAAG AAATTCCTCC TGACACATAA TAATCAGAAC AACAAATGCA CTAATAGATA 1000
     TACTTATATG TCCTTCGGAT GTCTTGAGGT TTATCTGACC TGGTCTTTTC TTTAAGGAGG ACTGTGTATT ATTAGTCTTG TTGTTTACGT GATTATCTAT

1001 GAATAGATAT AATAGATAGA ATATTAAAAG CAGTAAGGGA GAAAAGTCAA GTAACATATA AAGGCAGACC TACCAGAATT ACACCAGACT TTTCACCAGA 1100
     CTTATCTATA TTATCTATCT TATAATTTTC GTCATTCCCT CTTTTCAGTT CATTGTATAT TTCCGTCTGG ATGGTCTTAA TGTGGTCTGA AAAGTGGTCT
                                                                Hpa II
1101 GACAATGAAA GCCAGAAGAG CCTGGACAGA TGTTATACAG ACACTAAGAG AACACAAATG CCAGCCTAGG CTACTATGGC CAAACTCTCA ATTACCATAG 1200
     CTGTTACTTT CGGTCTTCTC GGACCTGTCT ACAATATGTC TGTGATTCTC TTGTGTTTAC GGTCGGATCC GATGATACCG GTTTGAGAGT TAATGGTATC

1201 ATGGAGAAAC CAAAGTATTC CACGACAAAA CCAAATTTAC ACATTATCTT TCCACGAATC CAGCCCTTCA AAGGATAATA ACAGAAAAAC AAACAAACAA 1300
     TACCTCTTTG GTTTCATAAG GTGCTGTTTT GGTTTAAATG TGTAATAGAA AGGTGCTTAG GTCGGGAAGT TTCCTATTAT TGTCTTTTTG TTTGTTTGTT

1301 ACAAACAAAC AAACAAAAAA ACAATACAAG GACGAAAATC ACTCCCTAGA AAAAGCAAGA AAGTAATCCC TCAACAAACC AAAAGAAGAC AGCCACAGAA 1400
     TGTTTGTTTG TTTGTTTTTT TGTTATGTTC CTGCTTTTAG TGAGGGATCT TTTTCGTTCT TTCATTAGGG AGTTGTTTGG TTTTCTTCTG TCGGTGTCTT

1401 CAGAATGCCA ACTCTAATAA CAAAAATAAA AGGAAGCAAC ATTTACTTTT CCTTAATATC TCTTAATATC AATGGACTCA ATTCCCCAAT AAAAAGACAT 1500
     GTCTTACGGT TGAGATTATT GTTTTTATTT TCCTTCGTTG TAAATGAAAA GGAATTATAG AGAATTATAG TTACCTGAGT TAAGGGGTTA TTTTTCTGTA
                Acc I          Ava II
1501 AGACTAACAG AACTGTAGAC ACAAACAGGA CCCAACATTC TGCTGCTTAC AGGAAACCCA TCTCAGGGAA AAAGACAGAA ACTTACCTCA GCGTGAAAGG 1600
     TCTGATTGTC TTGACATCTG TGTTTGTCCT GGGTTGTAAG ACGACGAATG TCCTTTGGGT AGAGTCCCTT TTTCTGTCTT TGAATGGAGT CGCACTTTCC
                                                      Taq I
1601 CTGGAAAACA ATTTTCCAAG CAAATGGTCT GAAGAAACAG GCTGGAGTAG CCATTCTAAT ATCGAATAAA ATTGACTTCC AACCCAAAGT CATCAAAAAA 1700
     GACCTTTTGT TAAAAGGTTC GTTTACCAGA CTTCTTTGTC CGACCTCATC GGTAAGATTA TAGCTTATTT TAACTGAAGG TTGGGTTTCA GTAGTTTTTT

1701 GGAAAATAGG GACACTTCAT ATTCATCAAA GTTAAAATCC TCCAAGAGGA ACTCACAATT CTGAATATCT ATGCTCCAAA TGCAAGGGCA GTCACATTCA 1800
     CCTTTTATCC CTGTGAAGTA TAAGTAGTTT CAATTTTAGG AGGTTCTCCT TGAGTGTTAA GACTTATAGA TACGAGGTTT ACGTTCCCGT CAGTGTAAGT

1801 TTAAAGACAC ATTAGTAAAG TTCAAAGCAC ACATTGTACC TCACACAATA ATAGTGGGAG ACTTCAACAC ACCACTTTCA TCAATGGACA GATCGTGGAA 1900
     AATTTCTGTG TAATCATTTC AAGTTTCGTG TGTAACATGG AGTGTGTTAT TATCACCCTC TGAAGTTGTG TGGTGAAAGT AGTTACCTGT CTAGCACCTT

1901 ACAGAAACTA AACAGGGACA CAATGAACCT AACAGAAGTT ATGAAACAAA TGGACTTAAC AGATATCTAC AGAACATTTT ATCCTTAAAC AAAAGGTTTT 2000
     TGTCTTTGAT TTGTCCCTGT GTTACTTGGA TTGTCTTCAA TACTTTGTTT ACCTGAATTG TCTATAGATG TCTTGTAAAA TAGGAATTTG TTTTCCAAAA
                Ava II
2001 ACCTTCTTCT CAGCACGGTC CAAAATTGAC CATATAATTG TTCACAAAAC AGGCCTCAAC AGATACAAAA ATACTGAAAT CGTCCCATGC ATCCTATTAG 2100
     TGGAAGAAGA GTCGTGCCAG GTTTTAACTG GTATATTAAC AAGTGTTTTG TCCGGAGTTG TCTATGTTTT TATGACTTTA GCAGGGTACG TAGGATAATC

2101 ACCACCATGG ACTAAGGCTG ATCTTCAATA ACAACATAAA TAATGGAAAG CCAACATTCA CGTGGAAACT GAACAACACT CTTCTCAATG AAACCTTGGT 2200
     TGGTGGTACC TGATTCCGAC TAGAAGTTAT TGTTGTATTT ATTACCTTTC GGTTGTAAGT GCACCTTTGA CTTGTTGTGA GAAGAGTTAC TTTGGAACCA

2201 CAAGGAAGTA ATAAAGAAAG AAATTAAAGA CTTTTTAGAG TTTAATGAAA ATGAAGCAGA TGCTGGCGAG GATGTGGAGA AAGAGGAACA CTCCTCCATT 2300
     GTTCCTTCAT TATTTCTTTC TTTAATTTCT GAAAAATCTC AAATTACTTT TACTTCGTCT ACGACCGCTC CTACACCTCT TTCTCCTTGT GAGGAGGTAA
                Hind III
2301 GTTGGTGGGG CG..AAGCTT  2320
     CAACCACCCC GC..TTCGAA
```

Figure 3.34 Strategy for sequencing subclone 14SS1

Only those sites used for labelling following primary restriction are shown. The details of this figure are as described for Figure 3.21. The arrows are numbered to serve as a reference for the table below, outlining the details of the sequencing.

Sequence run	Labelled restriction site	Radionucleotide used	Restriction enzyme second cut	Strand sequenced A or B
1	AvaI	γ^{32}P-ATP	XbaI	A
2	XbaI	α^{32}P-dCTP	AvaI	A
3	AccI	γ^{32}P-ATP	BglII	A
4	HindIII* (PstI (b))	α^{32}P-dCTP	AvaI	A
5	HindIII* (PstI (b))	γ^{32}P-ATP	EcoRI	A
6	HindIII* (PstI (c))	γ^{32}P-ATP	BglII	A
7	BglII	α^{32}P-dCTP	AvaI	A
8	EcoRI* (3'SmaI)	α^{32}P-dATP	PstI	A
9	HindIII* (5'SmaI)	α^{32}P-dCTP	XbaI	B
10	HindIII* (PstI (a))	α^{32}P-dCTP	XbaI	B
11	AvaI	α^{32}P-dCTP	PstI	B
12	XbaI	γ^{32}P-ATP	AvaI	B
13	XbaI	α^{32}P-dCTP	PstI	B
14	AccI	γ^{32}P-ATP	AvaI	B
15	EcoRI* (PstI (b))	γ^{32}P-ATP	XbaI	B
16	HindIII* (PstI (b))	α^{32}P-dCTP	EcoRI	B
17	AvaII (b)	γ^{32}P-ATP	XbaI	B
18	HindIII* (PstI (c))	γ^{32}P-ATP	EcoRI	B
19	AvaII (c)	α^{32}P-dCTP	BglII	B
20	BglII	γ^{32}P-ATP	XbaI	B
21	BglII	α^{32}P-dCTP	EcoRI	B

* Polylinker restriction site of pUC18

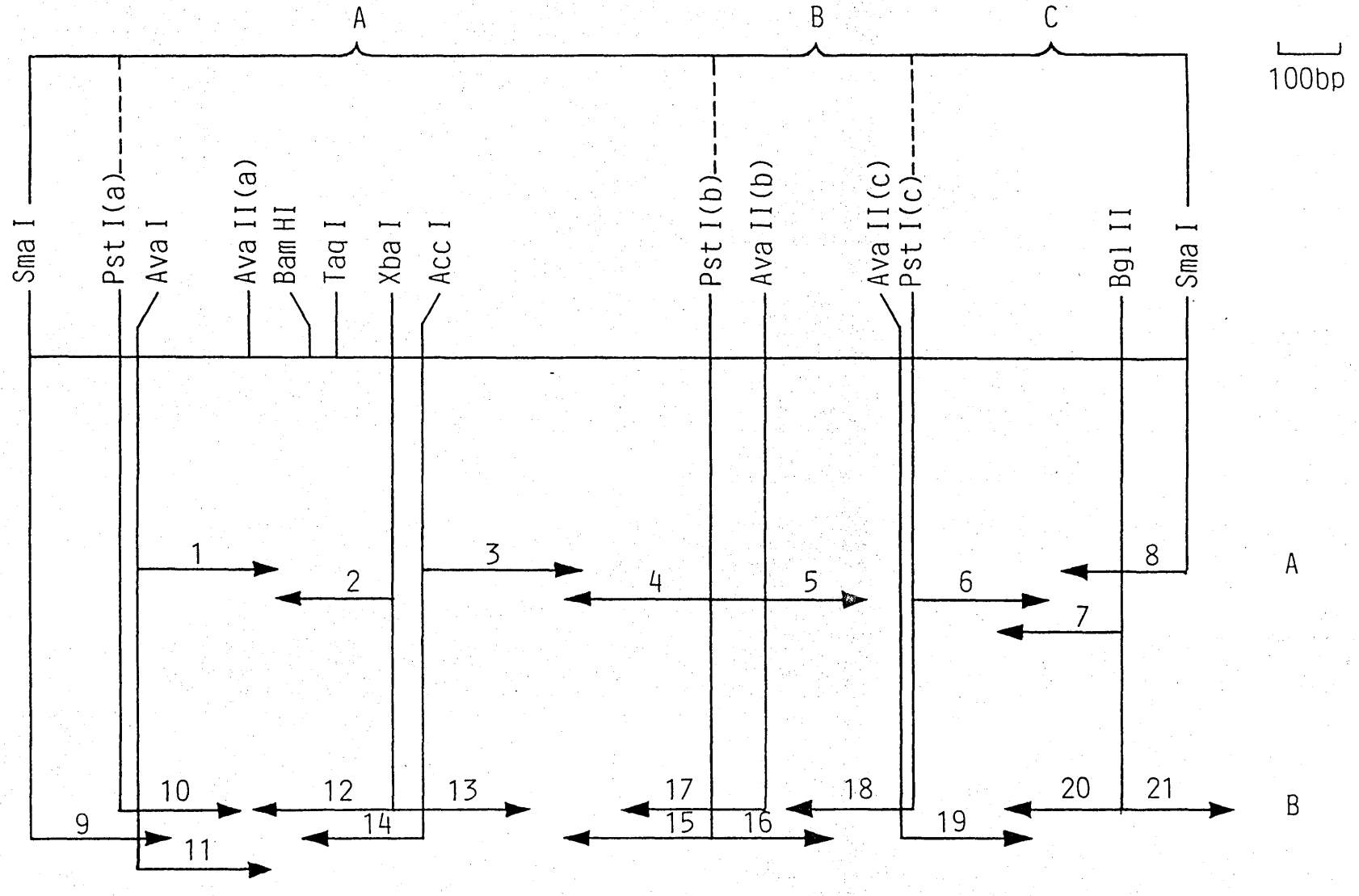

14SS1
A
B
C
100bp
Sma I
Pst I(a)
Ava I
Ava II(a)
Bam HI
Taq I
Xba I
Acc I
Pst I(b)
Ava II(b)
Ava II(c)
Pst I(c)
Bgl II
Sma I
1
2
3
4
5
6
7
8
A
9
10
11
12
13
14
15
16
17
18
19
20
21
B

Figure 3.35 Nucleotide sequence of subclone 14SS1

The nucleotide sequence was designated RH1 in the text as it contains a portion of the right-hand arm of the stem of the foldback structure in λmA14 (Figure 3.31). To assist subsequent comparison, this sequence is numbered in reverse, nucleotide 1 is the first base at the 3' SmaI site (Figure 3.34).

```
                Sma I
   1  CCCGGGCCTA AAGAAAGAAG AGATATGTGT CTAGGCCTAT TCCTGAACAT TGAAGAGGCC CGGACTAAAA GCAAAATAGT TGAGGGCTAG GGTCAAAAGC  100
      GGGCCCGGAT TTCTTTCTTC TCTATACACA GATCCGGATA AGGACTTGTA ACTTCTCCGG GCCTGATTTT CGTTTTATCA ACTCCCGATC CCAGTTTTCG
                                    Bgl II
 101  AAGAAGTGAG GGGGCCTAGG TCTATCCCAG ATCTTTGTTT GAATCCTAGC CTAAAGAAAG AATTGATGTG GGCCTAGGCC ATCCCTGACC CTTGAAGAGG  200
      TTCTTCACTC CCCCCGATCC AGATAGGGTC TAGAAACAAA CTTAGGATCG GATTTCTTTC TTAACTACAC CCGGATCCGG TAGGGACTGG GAACTTCTCC

 201  CCCTAGGCAA AGCAAGAAGT GAAAGTGCCT AGGTCTATAC CTGACTTTTG AAGAAGCCAG AGCCTAAAGA AAGAAGTGAT GTGGGTGTGG GTCTATTCCC  300
      GGGATCCGTT TCGTTCTTCA CTTTCACGGA TCCAGATATG GACTGAAAAC TTCTTCGGTC TCGGATTTCT TTCTTCACTA CACCCACACC CAGATAAGGG

 301  GATCACTTGA AGAGGCCTTG GCAATAAGCA AGAAATTTAA AGATGCCTAG GCCCAATGCA AGACGTGAAG AGGGCCTAGA CCTACACCTG ACCCTTGAAA  400
      CTAGTGAACT TCTCCGGAAC CGTTATTCGT TCTTTAAATT TCTACGGATC CGGGTTACGT TCTGCACTTC TCCCGGATCT GGATGTGGAC TGGGAACTTT
                                                                                  Pst I          Ava II
 401  GCTGCCTAGG CCTAAAGAAA GAAGTGCCCT TCTGGTCGGC ACCAGCACAG GGGCATCTTG GGCACAGAGT CTGCAGACAC CCCCAAGGTC CCCAGAGGAC  500
      CGACGGATCC GGATTTCTTT CTTCACGGGA AGACCAGCCG TGGTCGTGTC CCCGTAGAAC CCGTGTCTCA GACGTCTGTG GGGGTTCCAG GGGTCTCCTG

 501  TCTCCATGGG ATCTTAAGAC CTCTGGTGAG TGGAACACAA CTTCTGCTCC AATCCAATCG CATGGAACCT GAGACAGCAT GCTTAGGGAA GCAAGAAACC  600
      AGAGGTACCC TAGAATTCTG GAGACCACTC ACCTTGTGTT GAAGACGAGG TTAGGTTAGC GTACCTTGGA CTCTGTCGTA CGAATCCCTT CGTTCTTTGG

 601  TGGCCTGACA GGTCACAAGT CCTTTCTGGT AGGCACCAGC ACAGGGCACA TTGGGCTCAG AGTATGCGGA CATGCCCAAG GTTCCCAGAG GACTCTCCAC  700
      ACCGGACTGT CCAGTGTTCA GGAAAGACCA TCCGTGGTCG TGTCCCGTGT AACCCGAGTC TCATACGCCT GTACGGGTTC CAAGGGTCTC CTGAGAGGTG
          Ava II
 701  AGGATCTTGG GACCTCTGGG GAGTGGAACA CAACTTCTGC CAGGAGGCAG GTTCAAACAC CAGACATCTG GGCACCTTCC CTGCAAGAGG AGAGCTTGCC  800
      TCCTAGAACC CTGGAGACCC CTCACCTTGT GTTGAAGACG GTCCTCCGTC CAAGTTTGTG GTCTGTAGAC CCGTGGAAGG GACGTTCTCC TCTCGAACGG
 Pst I
 801  TGCAGAGAGT ACTCTGACCA CTGAAACTCA GGAGAAGCTA GTCTCCAGGT CTGCTGAAAG AGGCTAACAT AATCACTGGA GGAACAATCT CTAAACCGAG  900
      ACGTCTCTCA TGAGACTGGT GACTTTGAGT CCTCTTCGAT CAGAGGTCCA GACGACTTTC TCCGATTGTA TTAGTGACCT CCTTGTTAGA GATTTGGCTC

 901  ACAACTATAA CAACTAACTC CAGAGATTAC CAGATGGCTA AAGGCAAACG TAAGAATCTT ACTAACAGAA ACCAATACCA CTCACCATCA TCAGAATGAA 1000
      TGTTGATATT GTTGATTGAG GTCTCTAATG GTCTACCGAT TTCCGTTTGC ATTCTTAGAA TGATTGTCTT TGGTTATGGT GAGTGGTAGT AGTCTTACTT

1001  GCATTCCCAC CCCACCCAGT CTTGGGCACC CCAACACACT TGAAAAACAT CCCGGAATTA AAGCATATCT CATGATGATG GTAGAGGACA TCAAGAAGGA 1100
      CGTAAGGGTG GGGTGGGTCA GAACCCGTGG GGTTGTGTGA ACTTTTTGTA GGGCCTTAAT TTCGTATAGA GTACTACTAC CATCTCCTGT AGTTCTTCCT

1101  CTTTAACAAC TCAGTTAAAG AAATACAAGA GAAAATTGCT AAAGAGTTAC AAGTCCTTAA AGAAAACCAC GAAAACACAA CCAAACAGGA AGAAGTCCTT 1200
      GAAATTGTTG AGTCAATTTC TTTATGTTCT CTTTTAACGA TTTCTCAATG TTCAGGAATT TCTTTTGGTG CTTTTGTGTT GGTTTGTCCT TCTTCAGGAA
                                                                                  Acc I
1201  AAAGAAAAAC AGGAAAACAT ATCCAAACAG GTGATGGAAA TGAATAAAAC CATACTAGAC CTATAAAGGG AAGTAGACAC AATAAAGAAA ACCCAAAGTG 1300
      TTTCTTTTTG TCCTTTTGTA TAGGTTTGTC CACTACCTTT ACTTATTTTG GTATGATCTG GATATTTCCC TTCATCTGTG TTATTTCTTT TGGGTTTCAC
                Xba I
1301  AGGCAACACT GGAAATAGAA ACTCTAGAAA AGAAATCTGG AACCATAGAT GCAAGCATCA GCAACAGAAT ACAAGAAATG GAAGAGAGAA TCTCAGGTGC 1400
      TCCGTTGTGA CCTTTATCTT TGAGATCTTT TCTTTAGACC TTGGTATCTA CGTTCGTAGT CGTTGTCTTA TGTTCTTTAC CTTCTCTCTT AGAGTCCACG
                Taq I                                   Bam HI
1401  AGAAGATTCC ATAGAGAACA TCGACACAAC AGTCAAAGAA AATACAAAAT GCAAAAGGAT CCTAACTCAA AACATTCAGG TAATCCAGGA CACAATGAGA 1500
      TCTTCTAAGG TATCTCTTGT AGCTGTGTTG TCAGTTTCTT TTATGTTTTA CGTTTTCCTA GGATTGAGTT TTGTAAGTCC ATTAGGTCCT GTGTTACTCT

1501  AGACCAAACC TACGGATAAT AGGAATTGAT GAGAATGAAG ATTTTCAACT TAAAGGGCCA GCAAATATTT TCAACAAAAT AATAGAAGAA AACTTCCCAA 1600
      TCTGGTTTGG ATGCCTATTA TCCTTAACTA CTCTTACTTC TAAAAGTTGA ATTTCCCGGT CGTTTATAAA AGTTGTTTTA TTATCTTCTT TTGAAGGGTT
                                                  Ava II
1601  ACCTAAAGAG ATGCCCATGA ACATACAAGA AGCCTACAGA ACTCCAAATA GACTGGACCA GAAAAGAAAT TCCTTCTGAC ACATAATAAT CAGAACAACA 1700
      TGGATTTCTC TACGGGTACT TGTATGTTCT TCGGATGTCT TGAGGTTTAT CTGACCTGGT CTTTTCTTTA AGGAAGACTG TGTATTATTA GTCTTGTTGT
                          Ava I                                   Pst I
1701  AATGCACTAA ATAAAGATAG AATATTAAAA GCAGGGACCT CGGGAGCCAT CTTGGTTCTG GGACTCTGCA GAAAGTAGTC TGCACAGGTG AGAGTGTGCG 1800
      TTACGTGATT TATTTCTATC TTATAATTTT CGTCCCTGGA GCCCTCGGTA GAACCAAGAC CCTGAGACGT CTTTCATCAG ACGTGTCCAC TCTCACACGC

1801  AATTGCAGAA GCTAACAGCT TCTGGGGCGG CAAGAGCCAC AGAGTTTCTG GCAGCGCCAT TTTCAGGGCT CCAGACATCC GGCAACCTCT GGCCACCCAC 1900
      TTAACGTCTT CGATTGTCGA AGACCCCGCC GTTCTCGGTG TCTCAAAGAC CGTCGCGGTA AAAGTCCCGA GGTCTGTAGG CCGTTGGAGA CCGGTGGGTG
                Sma I
1901  AGGTGTATGC TTGGCCCGGG  1920
      TCCACATACG AACCGGGCCC
```

subclone 14HH4A, which contains the upper right-hand arm of the stem (b_R). This sequence was designated RH2 and is shown in Figure 3.37.

The electron micrograph regions which constitute the stem of the foldback structure were located by comparison of the nucleotide sequences described above, using a computer programme (PALIGN, section 2.2.18).

Figure 3.38 is a comparison of the left-hand arm of the stem DNA (LH) with the lower right-hand arm (RH1), the RH1 sequence being reversed.

Figure 3.39 is a comparison of the left-hand arm of the stem DNA (LH) with the upper right-hand arm (RH2), the RH2 sequence being reversed.

Figure 3.40 is a comparison of the lower and upper right-hand arm DNA sequences (RH1 and RH2), both sequences being reversed.

Comparison of the stem sequences indicated their relationship and is illustrated in Figure 3.41.

3.2.3 Stem DNA databank search

The EMBL database (Heidelberg, West Germany) was searched in order to try to determine the nature of the stem DNA. The first stem sequences obtained were from the subclone 14HH1, and included 550bp of sequence to the right of the 3' end of the actin region, (Figure 3.32). Comparison of this sequence with those in the EMBL and GenBank databases using the programme WORDSEARCH (section 2.2.18), revealed no other sequences with significant homology to it.

3.2.4 Stem DNA mouse genomic blot

Further analysis of the 'stem' DNA was therefore undertaken by hybridising a ^{32}P-labelled SstI-AvaII fragment of stem DNA, from the

Figure 3.36 Strategy for sequencing subclone 14HH4A

This subclone was partially sequenced from three restriction sites HindIII, AvaII and BglII and the sequence data from each has been designated (i), (ii) and (iii) respectively. The details of this figure are as described for Figure 3.21. The arrows are numbered to serve as a reference for the table below, outlining the details of the sequencing.

Sequence run	Labelled restriction site	Radionucleotide used	Restriction enzyme second cut	Strand sequenced A or B
1	HindIII	γ^{32}P-ATP	EcoRI	A
2	AvaII	γ^{32}P-ATP	EcoRI	A
3	EcoRI* (BglII)	α^{32}P-dATP	HindIII	A
4	HindIII	α^{32}P-dCTP	EcoRI	B
5	EcoRI* (BglII)	α^{32}P-dATP	HindIII	B

* Polylinker restriction site of pUC18

14HH4A
Hind III
Pst I
Ava II
Taq I
Ava II
Pst I
Bgl II
100bp
(i)
(ii)
(iii)
1
2
3
4
5
A
B

Figure 3.37 Partial nucleotide sequence of subclone 14HH4A

The nucleotide sequence was designated RH2 in the text as it contains a portion of the right-hand arm of the stem of the foldback structure in λmA14 (Figure 3.31). To assist subsequent comparison, this sequence is numbered in reverse, nucleotide 1 is the first base at the BglII site (Figure 3.36).

```
          BglII
(i)       AGATCT...C CTCGGTACCC GGGGATCTAC AAGTAGAAGT AGAAACAATA AAGAAAACCC AAAGGGAGAC AACTCTGGAA ATAGAAATCC TAGGAAAGAA   100
          TCTAGA...G GAGCCATGGG CCCCTAGATG TTCATCTTCA TCTTTGTTAT TTCTTTTGGG TTTCCCTCTG TTGAGACCTT TATCTTTAGG ATCCTTTCTT

          ATCAGGAAAC ATAGATGTGA GCATAGCAAG AGAATACAAG AGATGAAAGA AAAAAAATCT CAGGTGCAGA AGGTTACATA CAATCAAAAA AAATGCAAAA   200
          TAGTCCTTTG TATCTACACT CGTATCGTTC TCTTATGTTC TCTACTTTCT TTTTTTTAGA GTCCTCGTCT TCCAATGTAT GTTAGTTTTT TTTACGTTTT

          TGCAAAAAGG TTCCAACTCA AAACATTCAA GAAATTCAAG ACACAATGAT....                                                   250
          ACGTTTTTCC AAGGTTGAGT TTTGTAAGTT CTTTAAGTTC TGTGTTACTA....

(ii)      AAAGGTAATA AAGGGAAAAC TCCAACACAA AGAGGGAAAT TATGCCTTAG AAAAAGCAAG AAAGTAATCC TCCAACAAAC CTAAAAGAAG ATAGCCACAA   830
          TTTCCATTAT TTCCCTTTTG AGGTTGTGTT TCTCCCTTTA ATACGGAATC TTTTTCGTTC TTTCATTAGG AGGTTGTTTG GATTTTCTTC TATCGGTGTT

          GAACAGAATC CCAACTCTAA CAACAAAAAT AAGAGGACGC AATAACTACT TTCCTTAATA TCTCTTAATA TCAATAGACT CAATTCCCCA ATAATAGACA   930
          CTTGTCTTAG GGTTGAGATT GTTGTTTTTA TTCTCCTGCG TTATTGATGA AAGGAATTAT AGAGAATTAT AGTTATCTGA GTTAAGGGGT TATTATCTGT
                                Ava II
          TAAGAGACTG GCTAC..... ..GGACC...                                                                            967
          ATTCTCTGAC CGATG..... ..CCTGG...

(iii)     TAGATAAAGA CTCAAAATCT TTGATAAATG TGTGATAGAT TAGTATAGCT CCTCCTCTTA CTCTAGCTGC AGCACATTTA TACTTTTCTA GCTATTCAGA  1590
          ATCTATTTCT GAGTTTTAGA AACTATTTAC ACACTATCTA ATCATATCGA GGAGGAGAAT GAGATCGACG TCGTGTAAAT ATGAAAAGAT CGATAAGTCT

          GATTTTTCCA CCCATCTTCT CAGACTTCTC AAGGTAAAAG TTATGGCAAG GGCTGTCTAC TTCACGTCAG GAATTGCAAC GAGCATATAC AGAGACGCAG  1690
          CTAAAAAGGT GGGTAGAAGA GTCTGAAGAG TTCCATTTTC AATACCGTTC CCGACAGATG AAGTGCTGTC CTTAACGTTG CTCGTATATG TCTCTGCGTC
             HindIII
          ...TTCGAA                                                                                                   1700
          ...AAGCTT
```

Figure 3.38 Comparison of nucleotide sequences ; LH and RH1

The nucleotide sequences of LH and RH1 are numbered as in Figures 3.33 and 3.35, respectively.

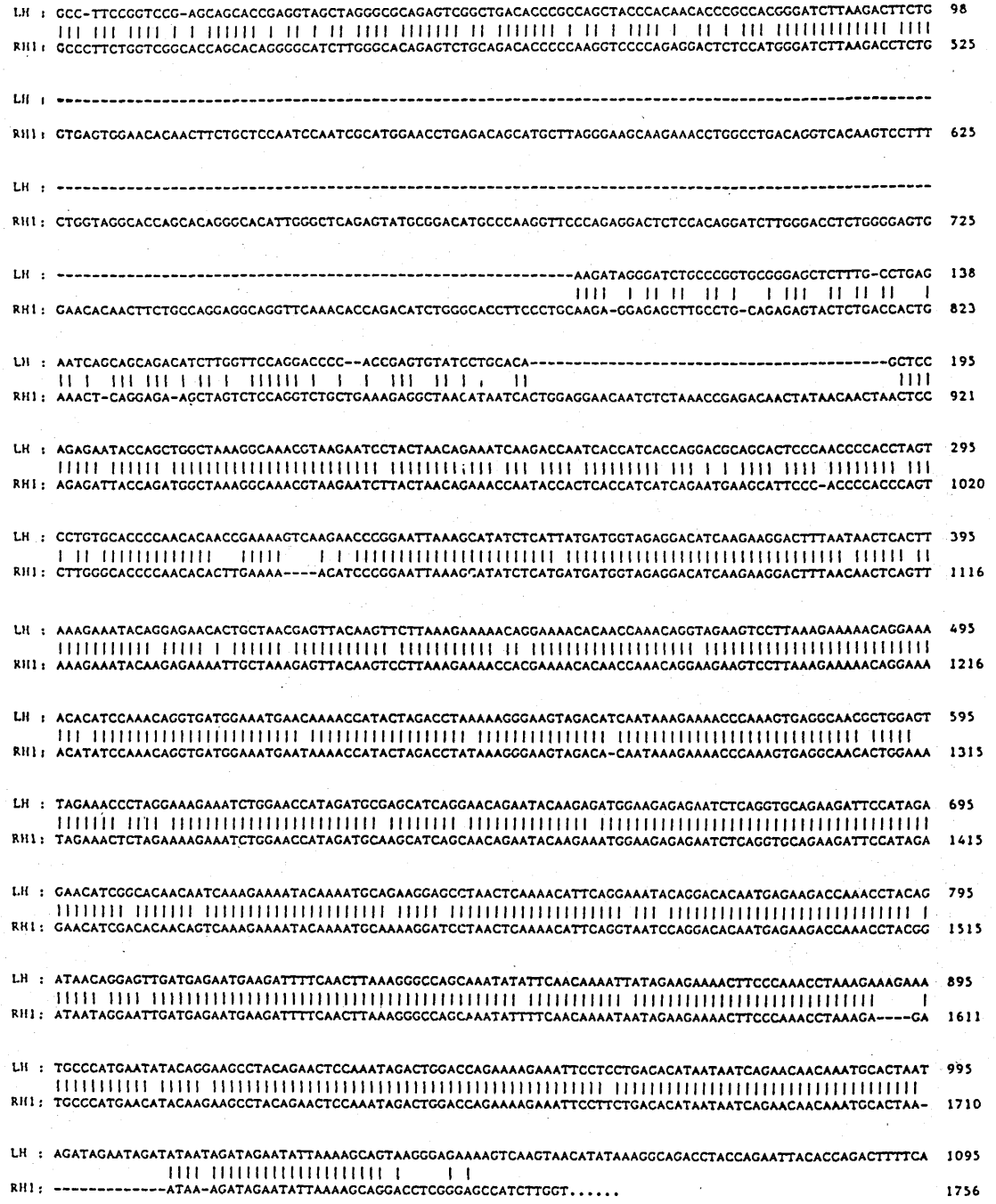


Figure 3.39 Comparison of nucleotide sequences : LH and RH2

The nucleotide sequences of LH and RH2 are numbered as in Figures 3.33 and 3.37 respectively.

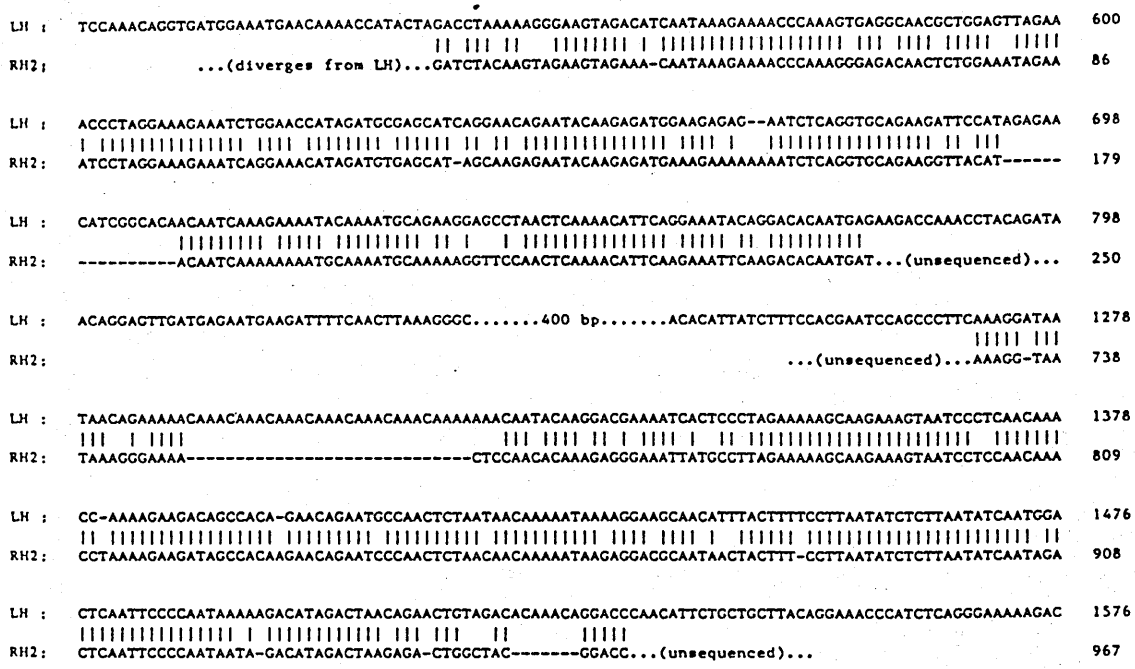


Figure 3.40 Comparison of nucleotide sequences : RH1 and RH2

The nucleotide sequences of RH1 and RH2 are numbered as in Figures 3.35 and 3.37, respectively.

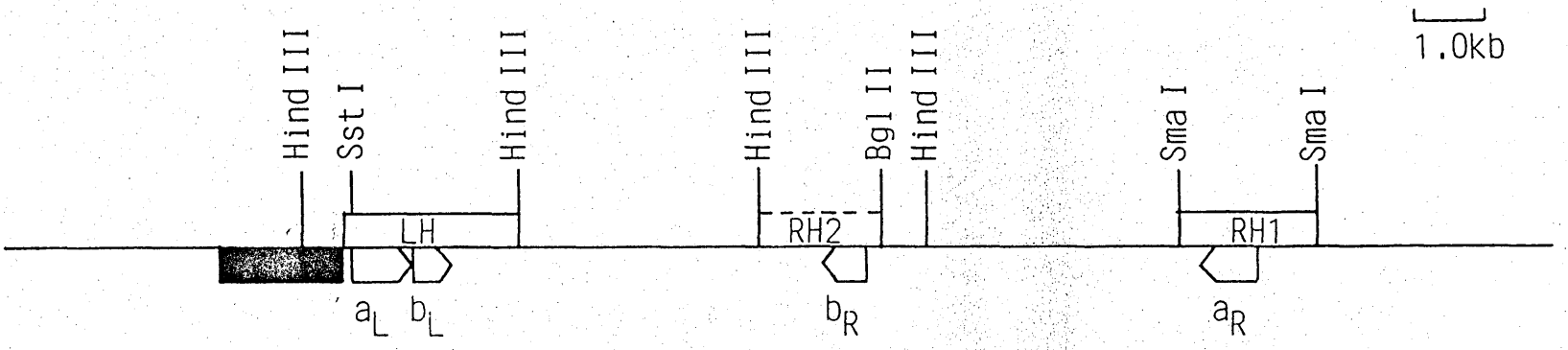
```
RH1:  GATGGAAATGAATAAAACCATACTAGACCTATAAAGGGAAGTAGACACAATAAAGAAAACCCAAAGTGAGGCAACACTGGAAATAGAAACTCTAGAAAAG  1332
RH2:        (Sequence diverges from RH1)...GAAGTAGAAACAATAAAGAAAACCCAAAGGGAGACAACTCTGGAAATAGAAATCCTAGGAAAG    98

RH1:  AAATCTGGAACCATAGATGCAAGCATCAGCAACAGAATACAAGAAATGGAAGAGAGA--ATCTCAGGTGCAGAAGATTCCATAGAGAACATCGACACAAC  1430
RH2:  AAATCAGGAAACATAGATGTGAGCAT-AGCAAGAGAATACAAGAGATGAAAGAAAAAAAATCTCAGGTGCAGAAGGTTACAT----------------AC   181

RH1:  AGTCAAAGAAAATACAAAATGCAAAA-GGATCCTAACTCAAAACATTCAGGTAATCCAGGACACAATGAG-AAGACCAAACCTACGGATAATAGGAATTGA  1529
RH2:  AATCAAAAAAAATGCAAAATGCAAAAAGG-TTCCAACTCAAAACATTCAAGAAATTCAAGACACAATGAT.....(unsequenced)...             250
```

Figure 3.41 Diagrammatic representation of the relationship between the various nucleotide sequences constituting the stem of the foldback structure within λmA14

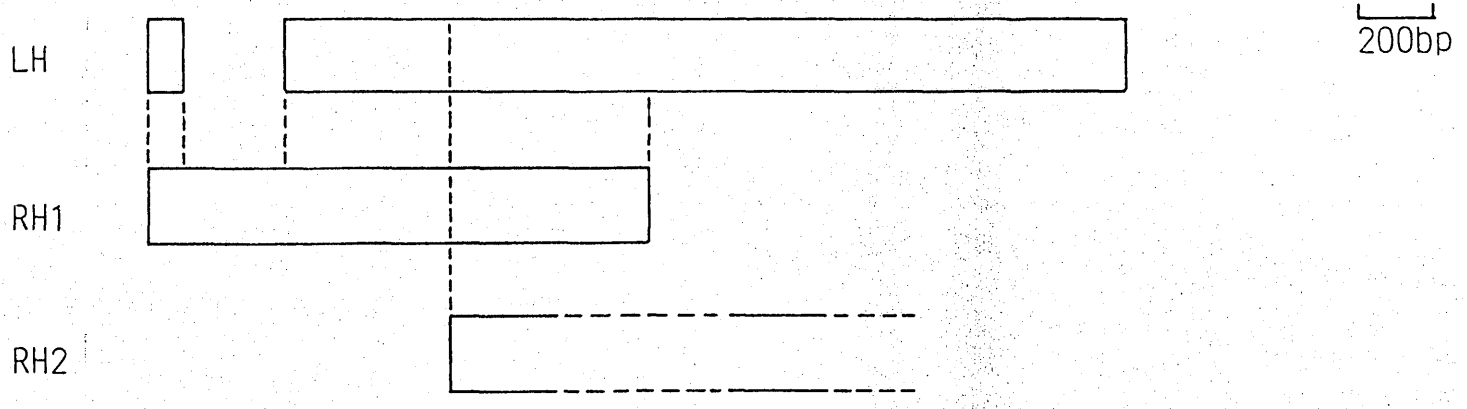
(i) The location of the nucleotide sequence LH, RH1 and RH2 within λmA14.

(ii) The arrangement of LH, RH1 and RH2 with homologous regions aligned.

(a) λmA14


Hind III
Sst I
Hind III
Hind III
Bgl II
Hind III
Sma I
Sma I
1.0kb
LH
RH2
RH1
a_L
b_L
b_R
a_R

(b)


LH
RH1
RH2
200bp

subclone 14HH1B (shown in Figure 3.26), against mouse DNA digested with BamHI and EcoRI. The results of the hybridisation, shown in Figure 3.42, indicated that the probe DNA was highly repetitive in the mouse genome. Also discrete bands were observed against a background smear, for example a 4.0kb BamHI fragment and a 3.0kb EcoRI fragment. The lengths of these fragments was similar to those in the previously characterised L1Md, mouse repetitive family (Fanning, 1983). This L1Md family is composed of members up to approximately 7.0kb in length, although the parts of it that had been characterised were the more abundant truncated members derived from the 3' end. The failure of the databank search to detect homologous sequences to the λmA14 stem DNA suggested that, if this were part of the L1Md family, it might be from the 5' end.

3.2.5 Comparison of the stem DNA of λmA14 with L1Md DNA sequence

At the beginning of 1986 the sequence of the first apparently 'full-length' L1Md member, L1Md-A2, was published (Loeb *et al.*, 1986), and this allowed comparison to be made with the λmA14 stem sequence.

Figure 3.43 shows that most of the LH sequence and specific regions within RH1 and RH2 sequences are homologous to L1Md DNA. Figure 3.44 is a diagrammatic representation of L1Md sequence within λmA14 and its relationship to the stem regions.

3.2.6 Location of extreme 3' end of λmA14 actin pseudogene

The L1Md DNA within LH extended in the leftward direction to the

Figure 3.42 Analysis of mouse genomic sequences homologous to the stem of the foldback structure within λmA14

(a) BALB/c mouse DNA was isolated as described in section 2.3.5, 10ug was digested with BamHI and EcoRI and subjected to gel electrophoresis through 0.7% agarose (lanes 1 and 2).

(b) The DNA was transferred to nitrocellulose and hybridised to ^{32}P-labelled SstI-AvaII restriction fragment from the subclone 14HH1B (Figure 3.26).

DNA marker 1 is λcI_{857} digested with HindIII and DNA marker 2 is pmS4 digested with TaqI (section 2.2.10).

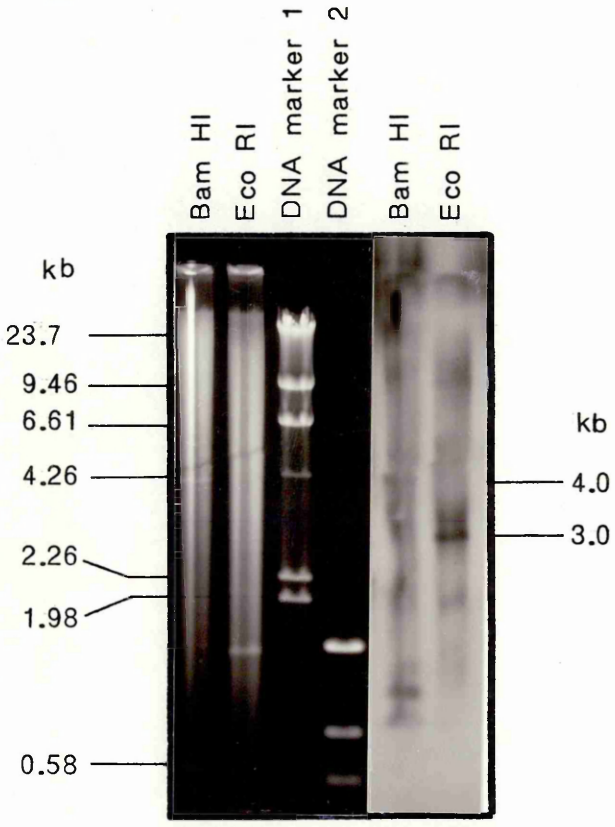
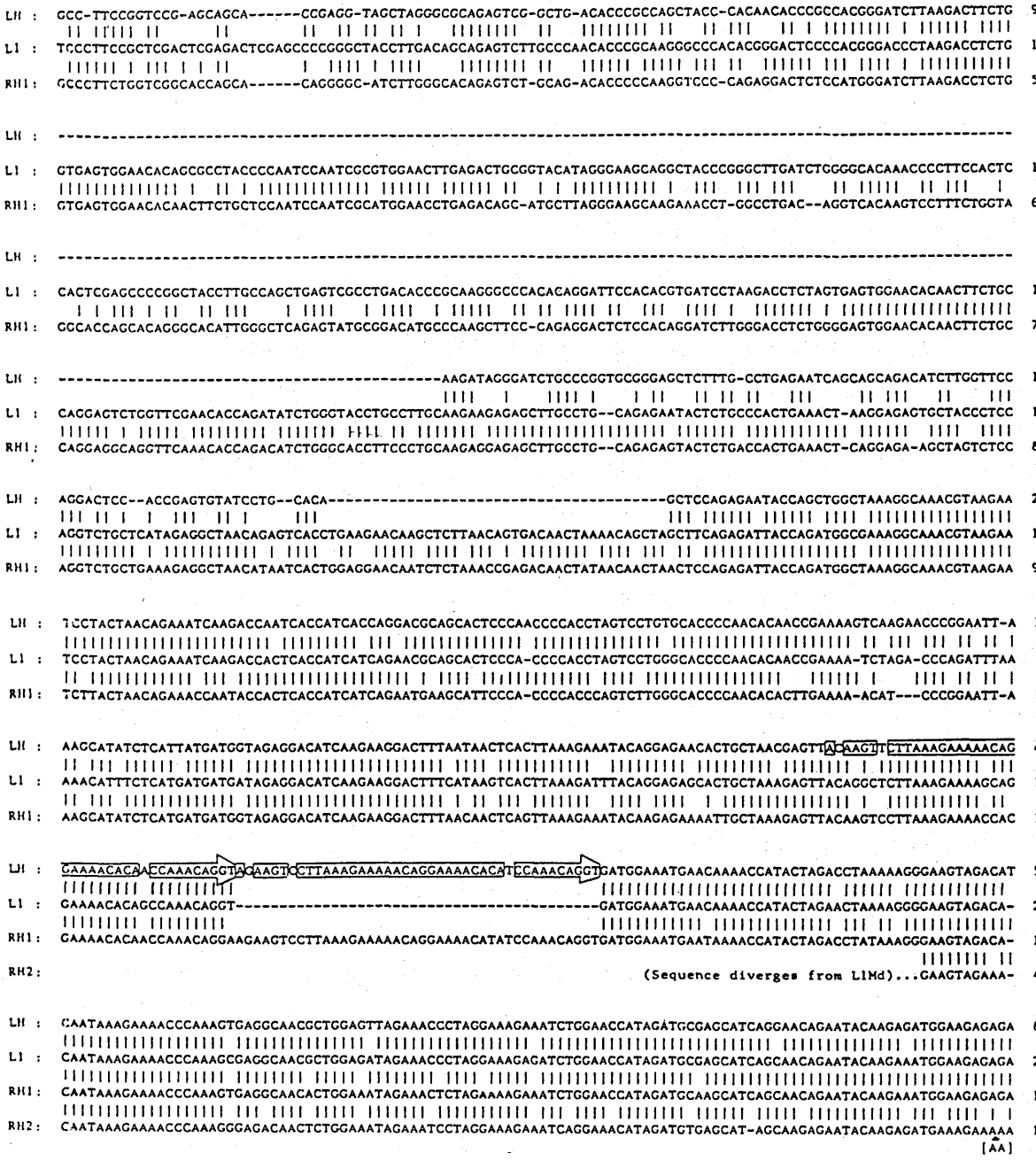
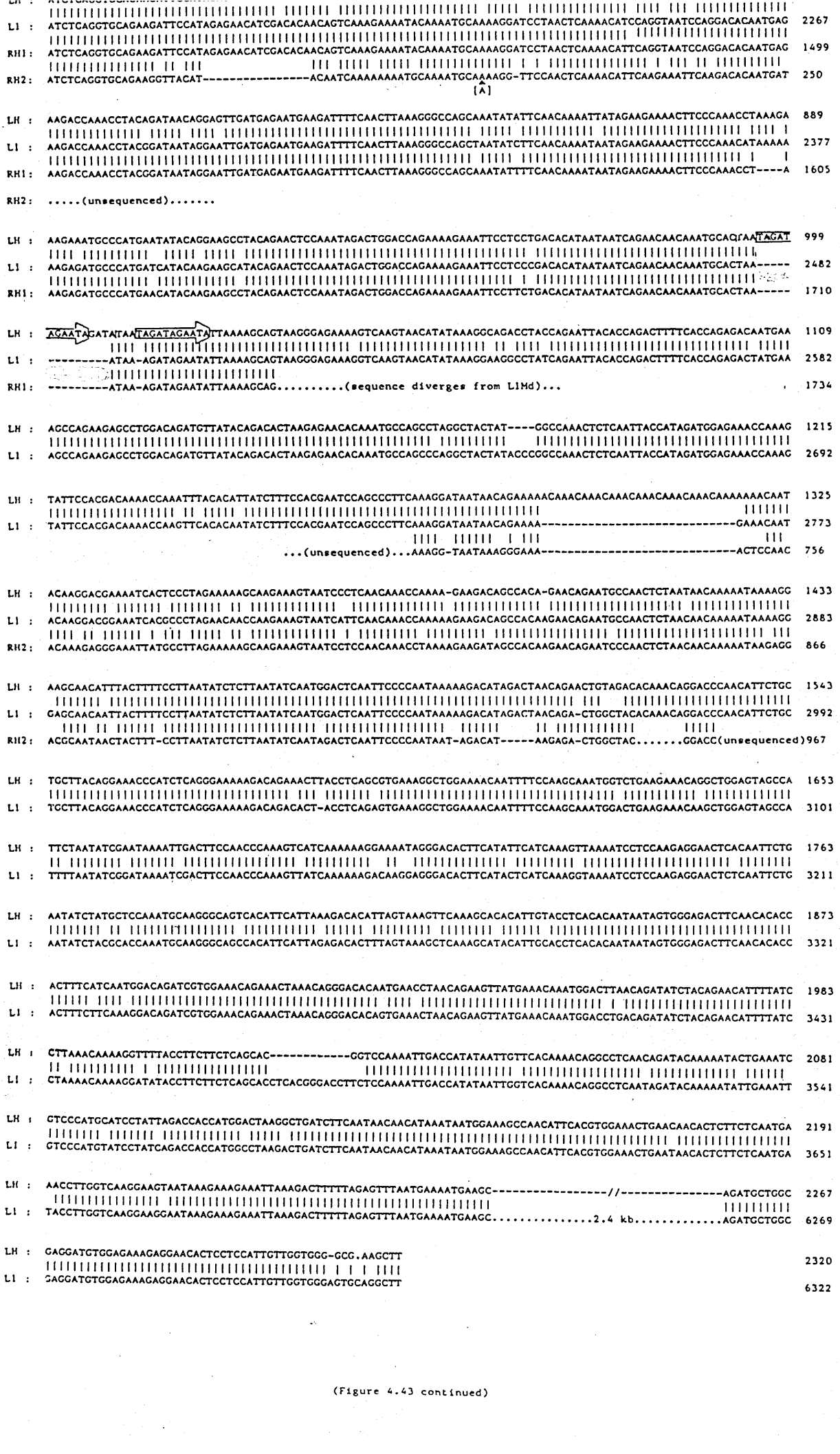

Bam HI
Eco RI
DNA marker 1
DNA marker 2
Bam HI
Eco RI
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
kb
4.0
3.0

Figure 3.43 Comparison of LH, RH1 and RH2 nucleotide sequences with a mouse repetitive DNA member L1Md-A2

The nucleotide sequence LH, RH1 and RH2 are numbered as in Figures 3.33, 3.35 and 3.37, respectively. The L1Md-A2 nucleotide sequence (designated L1) is numbered according to Loeb *etal.*, (1986).

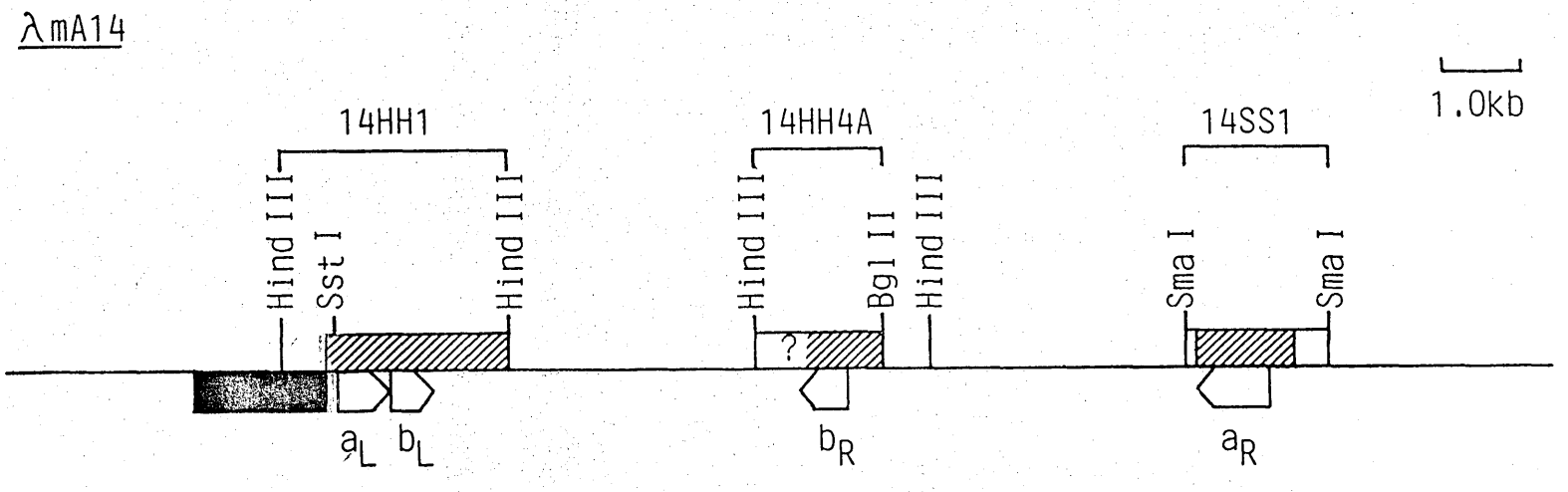




(Figure 4.43 continued)

Figure 3.44 Diagrammatic representation of the L1Md nucleotide sequence within λmA14 and its relationship to the stem regions

The λmA14 subclones which contain the electron micrograph stem sections are indicated. The stem sections are designated a and b, and are followed by a subscript L or R which respectively refers to the left and right-hand side of the stem. The L1Md nucleotide sequence within λmA14 is represented as barred regions.


λmA14
1.0kb
14HH1
14HH4A
14SS1
Hind III
Sst I
Hind III
Hind III
Bgl II
Hind III
Sma I
Sma I
?
a_L
b_L
b_R
a_R

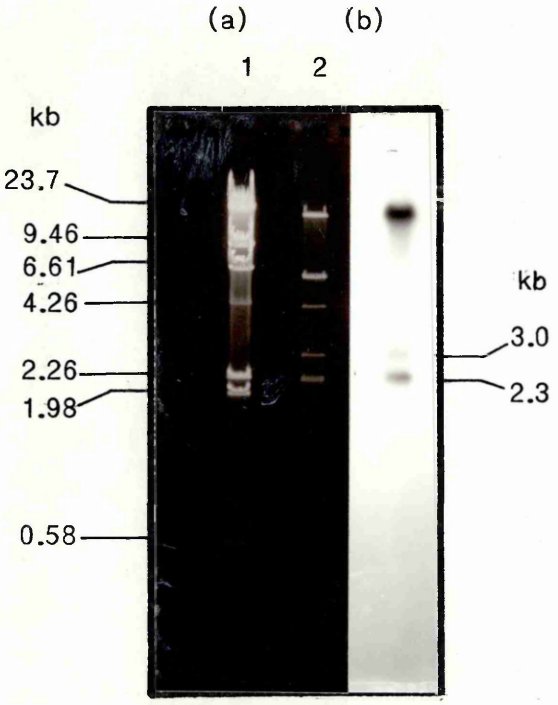
apparently truncated 3' end of the actin pseudogene as shown in Figure 3.44. It therefore appeared possible that the L1Md had inserted approximately 100bp from the expected 3' end of the actin pseudogene in λmA14, in which case the extreme 3' end of the pseudogene might be at the other extremity (3' end) of this L1Md sequence. To try to locate this 3' actin DNA a ^{32}P-labelled 150bp TaqI-PstI fragment from the actin pseudogene of λmA19, (Figure 2.3) was used as the probe. This contains 150bp of the 3' non-coding actin-like DNA (including the missing 100bp). The probe hybridised to the 2.3 and 3.0kb HindIII fragments of λmA14 (Figure 3.45). The 3.0kb fragment was that cloned as 14HH1, and contains the truncated 3' end of the actin pseudogene. The 2.3kb fragment corresponds to that cloned as 14HH2, and the hybridisation suggested that this contained the displaced 100bp 3' non-coding actin-like sequence. Further analysis indicated that the region in 14HH2 hybridising to the probe was more specifically positioned within the subclone 14HH2B (Figure 3.8). If it is assumed that the LH L1Md sequence in λmA14 continues uninterrupted from the right-hand HindIII site at the extremity of the 14HH1 where sequencing ended, one would predict on the basis of the results of Loeb *et al.*, (1986) that the 3' end of the L1Md sequence would be approximately 1.1kb further to the right. As 14HH2B is estimated to be 0.8kb to the right of 14HH1, (Figure 3.8), the 3' end of the LH L1Md sequence would be expected to be in 14HH2B. Thus the most reasonable interpretation of these results is that the remaining 100bp of the actin pseudogene would be at the 3' end of the L1Md sequence.

Figure 3.45 Location of extreme 3'end of the actin pseudogene by hybridisation of TaqI-PstI fragment from subclone MγA-ψ1, against digested λmA14

The ^{32}P-labelled TaqI-PstI restriction fragment was isolated from MγA-ψ1 (Figure 2.3), a subclone of λmA19 (Leader *et al.*, 1985). The 150bp fragment which contains the extreme 3' end of the non-coding DNA of the actin processed pseudogene in λmA19 was hybridised against digested λmA14.

(a) Photograph of the stained DNA gel, lane 1 is λcI_{857} digested with HindIII (section 2.2.10) and lane 2 is λmA14 digested with HindIII.

(b) Autoradiograph of the nitrocellulose, which corresponds to lane 2


(a)
(b)
1
2
kb
23.7
9.46
6.61
4.26
2.26
1.98
0.58
kb
3.0
2.3

CHAPTER 4 Discussion

4.1 Actin processed pseudogenes in λmA14 and λmA36

Although the major concern of this thesis is the DNA associated with the actin-like genes of clones λmA14 and λmA36, it is appropriate to begin this Discussion with a consideration of the results of partial nucleotide sequence determination on the actin-like genes themselves.

The portion of the actin-like nucleotide sequence in λmA14 includes bases number 1063 to 1998 of Figure 3.22. This is related to the coding sequence of an actin-like gene from amino-acid 1 to amino-acid 302, as shown in Figures 4.1 and 4.2. Of the 22 residues unique to cytoplasmic actins (Table 1.1), all were found in the predicted sequence of λmA14, (indicated by the underlined residues in Figures 4.1 and 4.2), except for the last two residues at positions 357 and 364, which are located outwith the sequence determined. There are 4 amino acids at the N-terminal end of the sequence which differentiate the cytoplasmic actin β and γ isoforms (Vanderckhove and Weber, 1979a). These are the amino acids at position 2 (β = Asp, γ =Glu), position 3 (β = Asp, γ = Glu), position 4 (β =Asp, γ = Glu) and position 10 (β = Val, γ = Ile). The sequence in λmA14 corresponds to Glu^2, Glu^3, Glu^4 and Ile^{10}, (Figure 4.1) and identified the cytoplasmic actin-like DNA in λmA14 as being related to the γ isoform. The actin-like region in λmA36 was only partially sequenced (Figure 3.24, parts (iii) and (iv)). The regions sequenced correspond to Pro^{70} to Asp^{157} and Cys ^{256} to Thr ^{302} (Figure 4.3). Within

Figure 4.1 Comparison of the predicted amino acid sequence (residues 1 - 50) for γ-actin with the corresponding region of λmA14

The predicted γ-actin amino acid sequence (Vandekerckhove & Weber, 1979a) is compared with the predicted amino acid sequence of the actin-like region in λmA14. The underlined residues are those unique to cytoplasmic actin and those labelled with (*) identify the cytoplasmic actin to be of the gamma (γ) type. The nucleotide differences in λmA14 which produce a residue alteration from the γ-actin amino acid sequence are indicated.

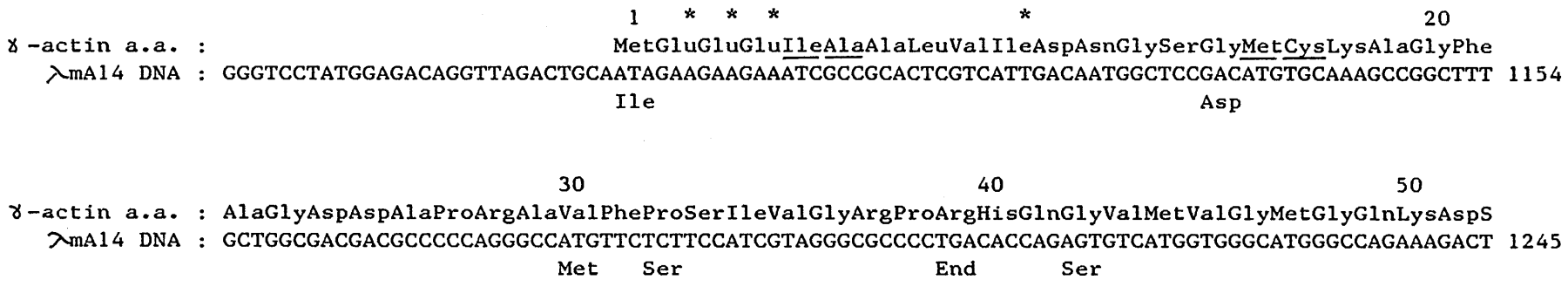
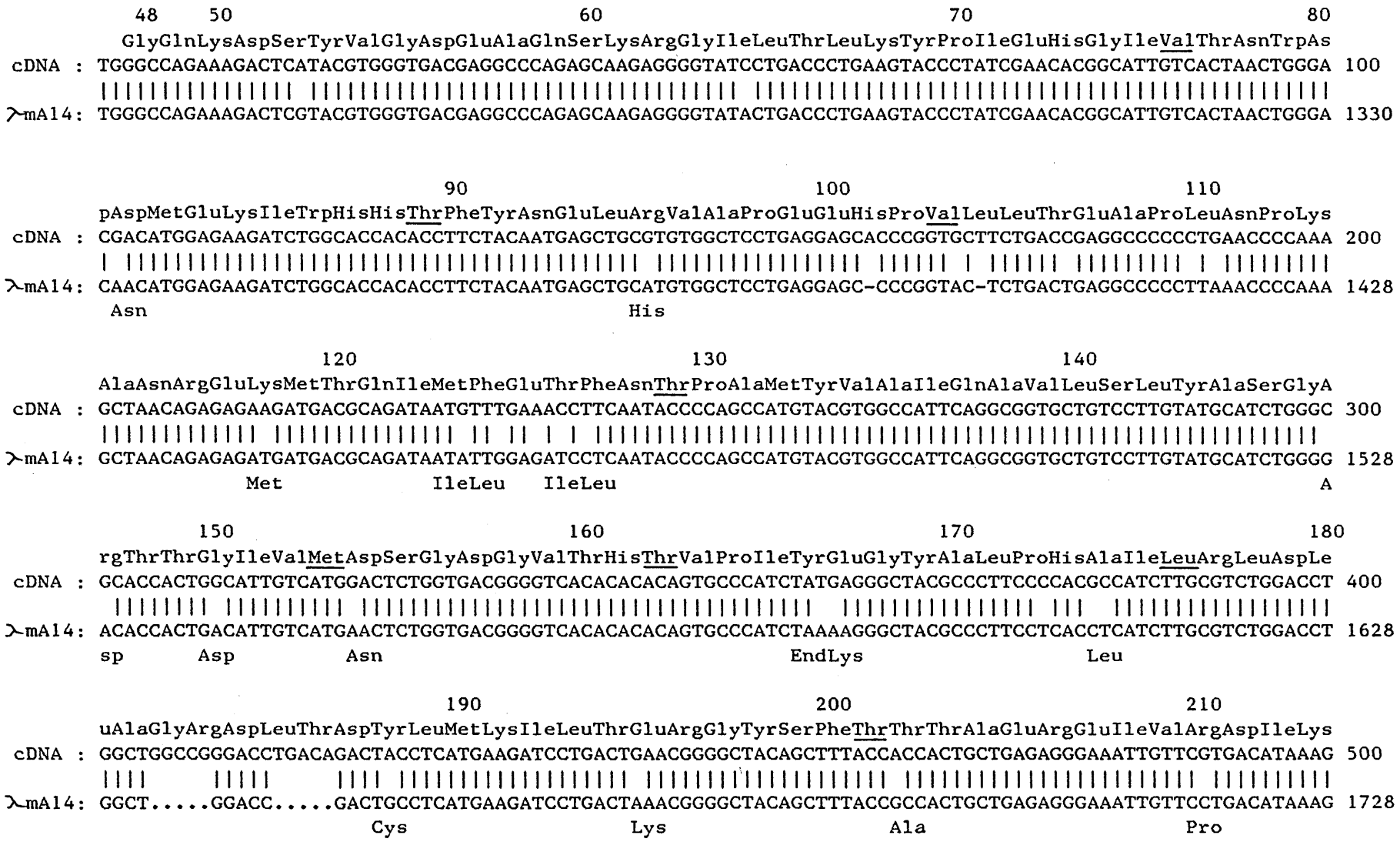
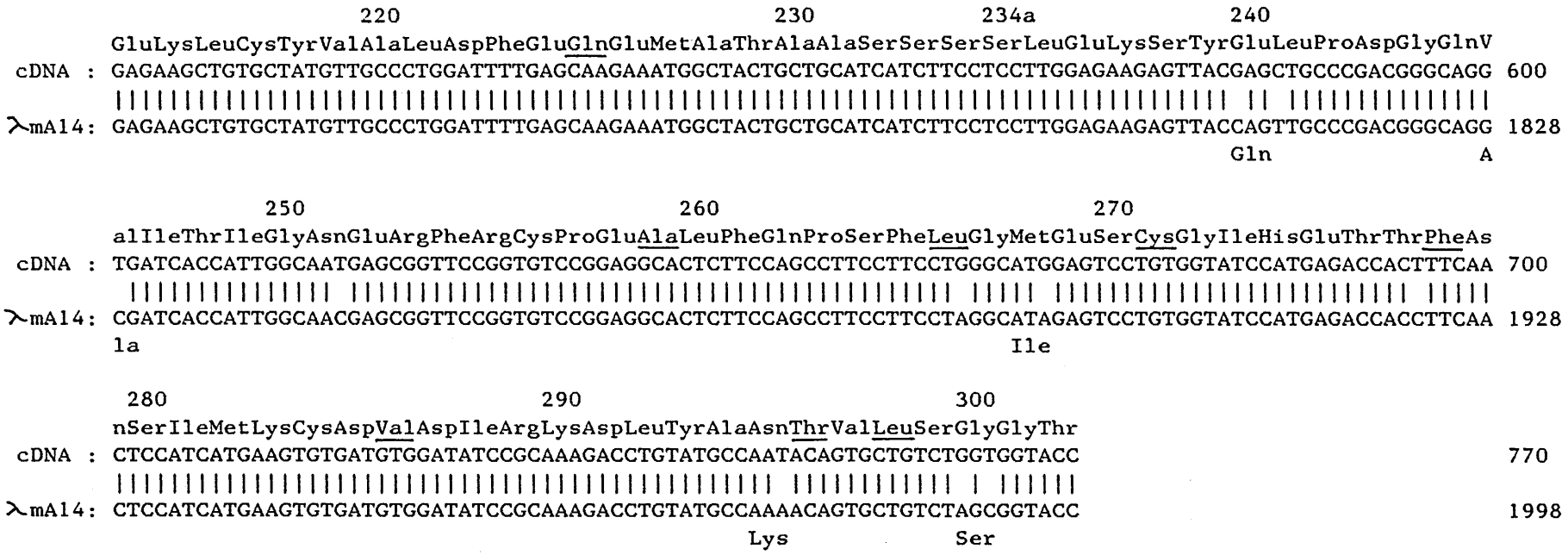

1 * * * * 20
γ-actin a.a. : MetGluGluGluIleAlaAlaLeuValIleAspAsnGlySerGlyMetCysLysAlaGlyPhe
λmA14 DNA : GGGTCCTATGGAGACAGGTTAGACTGCAATAGAAGAAGAAATCGCCGCACTCGTCATTGACAATGGCTCCGACATGTGCAAAGCCGGCTTT 1154
Ile Asp
30 40 50
γ-actin a.a. : AlaGlyAspAspAlaProArgAlaValPheProSerIleValGlyArgProArgHisGlnGlyValMetValGlyMetGlyGlnLysAspS
λmA14 DNA : GCTGGCGACGACGCCCCCAGGGCCATGTTCTCTTCCATCGTAGGGCGCCCCTGACACCAGAGTGTCATGGTGGGCATGGGCCAGAAAGACT 1245
Met Ser End Ser

Figure 4.2 Comparison of the nucleotide sequence (residues 48 - 302) of mouse γ-actin cDNA with the corresponding region of λmA14

The partial nucleotide sequence of the pseudo-coding region of a mouse γ-actin cDNA (Peter & Leader, unpublished) is compared with the corresponding actin-like region of λmA14. The underlined residues are those unique to cytoplasmic actin. The nucleotide differences in λmA14 which produce a residue alteration from the γ-actin amino acid sequence are indicated. Unsequenced regions are indicated by dots (·····) and nucleotide base deletions are indicated by a dash (-).

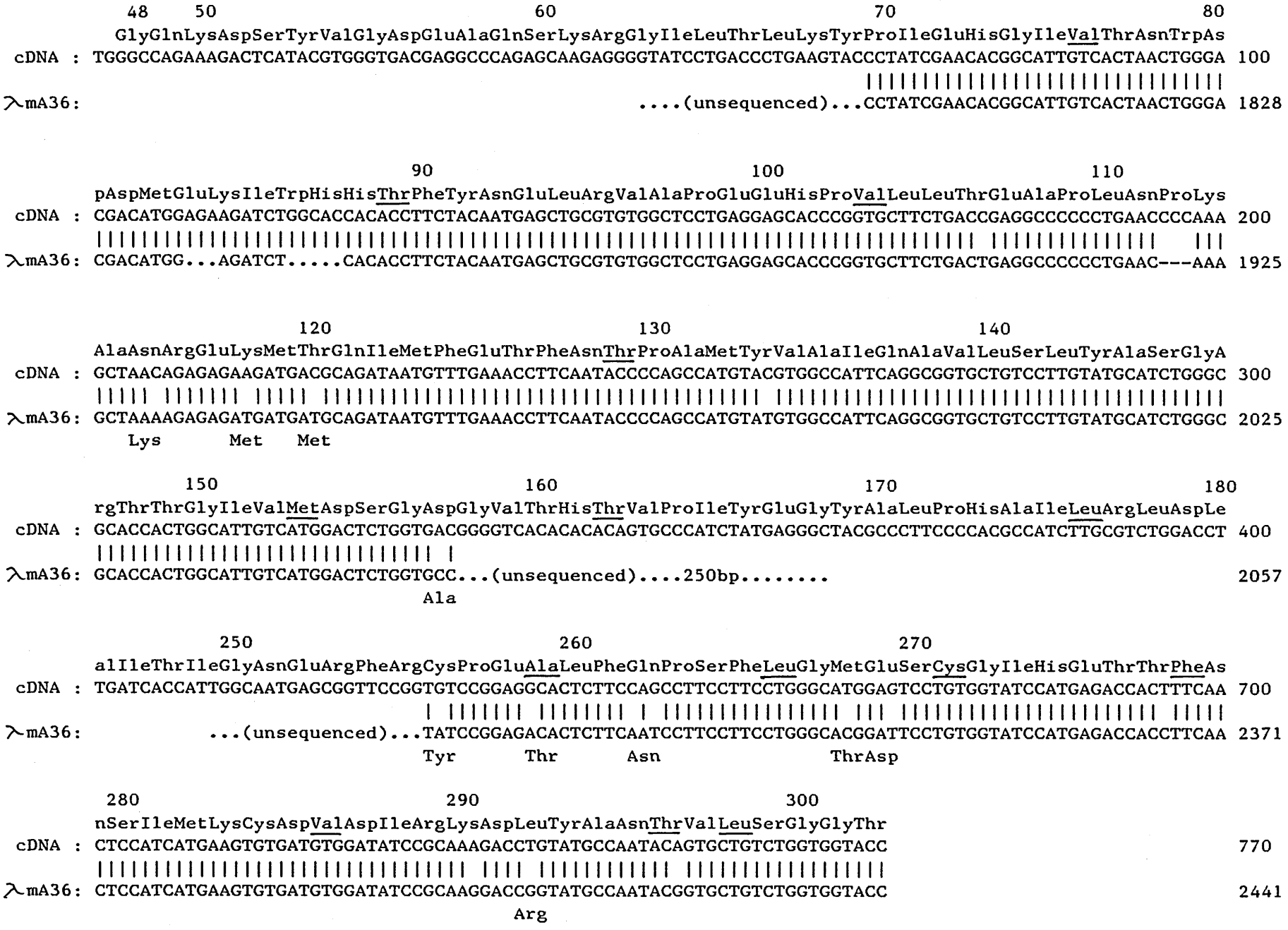

48 50 60 70 80
GlyGlnLysAspSerTyrValGlyAspGluAlaGlnSerLysArgGlyIleLeuThrLeuLysTyrProIleGluHisGlyIleValThrAsnTrpAs
cDNA : TGGGCCAGAAAGACTCATACGTGGGTGACGAGGCCCAGAGCAAGAGGGGTATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGA 100
λmA14: TGGGCCAGAAAGACTCGTACGTGGGTGACGAGGCCCAGAGCAAGAGGGGTATACTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGA 1330
90 100 110
pAspMetGluLysIleTrpHisHisThrPheTyrAsnGluLeuArgValAlaProGluGluHisProValLeuLeuThrGluAlaProLeuAsnProLys
cDNA : CGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGGCTCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAA 200
λmA14: CAACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCATGTGGCTCCTGAGGAGC-CCCGGTAC-TCTGACTGAGGCCCCCTTAAACCCCAAA 1428
Asn His
120 130 140
AlaAsnArgGluLysMetThrGlnIleMetPheGluThrPheAsnThrProAlaMetTyrValAlaIleGlnAlaValLeuSerLeuTyrAlaSerGlyA
cDNA : GCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCTTCAATACCCCAGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGC 300
λmA14: GCTAACAGAGAGATGATGACGCAGATAATATTGGAGATCCTCAATACCCCAGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGG 1528
Met IleLeu IleLeu A
150 160 170 180
rgThrThrGlyIleValMetAspSerGlyAspGlyValThrHisThrValProIleTyrGluGlyTyrAlaLeuProHisAlaIleLeuArgLeuAspLe
cDNA : GCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTGCCCATCTATGAGGGCTACGCCCTTCCCCACGCCATCTTGCGTCTGGACCT 400
λmA14: ACACCACTGACATTGTCATGAACTCTGGTGACGGGGTCACACACACAGTGCCCATCTAAAAGGGCTACGCCCTTCCTCACCTCATCTTGCGTCTGGACCT 1628
sp Asp Asn EndLys Leu
190 200 210
uAlaGlyArgAspLeuThrAspTyrLeuMetLysIleLeuThrGluArgGlyTyrSerPheThrThrThrAlaGluArgGluIleValArgAspIleLys
cDNA : GGCTGGCCGGGACCTGACAGACTACCTCATGAAGATCCTGACTGAACGGGGCTACAGCTTTACCACCACTGCTGAGAGGGAAATTGTTCGTGACATAAAG 500
λmA14: GGCT.....GGACC.....GACTGCCTCATGAAGATCCTGACTAAACGGGGCTACAGCTTTACCGCCACTGCTGAGAGGGAAATTGTTCCTGACATAAAG 1728
Cys Lys Ala Pro



(Figure 4.2 continued)

Figure 4.3 Comparison of the nucleotide sequence (residues 70 to 157 and 256 to 302) of mouse γ-actin cDNA with the corresponding region of λmA36

The partial nucleotide sequence of the pseudo-coding region of a mouse γ-actin cDNA (Peter & Leader, unpublished) is compared with the corresponding actin-like region of λmA36. The underlined residues are those unique to cytoplasmic actin. The nucleotide differences in λmA36 which produce a residue alteration from the γ-actin amino acid sequence are indicated. Unsequenced regions are indicated by dots (·····) and nucleotide base deletions are indicated by a dash (-).


48 50 60 70 80
GlyGlnLysAspSerTyrValGlyAspGluAlaGlnSerLysArgGlyIleLeuThrLeuLysTyrProIleGluHisGlyIleValThrAsnTrpAs
cDNA : TGGGCCAGAAAGACTCATACGTGGGTGACGAGGCCCAGAGCAAGAGGGGTATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGA 100
λmA36: ....(unsequenced)...CCTATCGAACACGGCATTGTCACTAACTGGGA 1828
90 100 110
pAspMetGluLysIleTrpHisHisThrPheTyrAsnGluLeuArgValAlaProGluGluHisProValLeuLeuThrGluAlaProLeuAsnProLys
cDNA : CGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGGCTCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAA 200
λmA36: CGACATGG...AGATCT.....CACACCTTCTACAATGAGCTGCGTGTGGCTCCTGAGGAGCACCCGGTGCTTCTGACTGAGGCCCCCCTGAAC---AAA 1925
120 130 140
AlaAsnArgGluLysMetThrGlnIleMetPheGluThrPheAsnThrProAlaMetTyrValAlaIleGlnAlaValLeuSerLeuTyrAlaSerGlyA
cDNA : GCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCTTCAATACCCCAGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGC 300
λmA36: GCTAAAAGAGAGATGATGATGCAGATAATGTTTGAAACCTTCAATACCCCAGCCATGTATGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGC 2025
Lys Met Met
150 160 170 180
rgThrThrGlyIleValMetAspSerGlyAspGlyValThrHisThrValProIleTyrGluGlyTyrAlaLeuProHisAlaIleLeuArgLeuAspLe
cDNA : GCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTGCCCATCTATGAGGGCTACGCCCTTCCCCACGCCATCTTGCGTCTGGACCT 400
λmA36: GCACCACTGGCATTGTCATGGACTCTGGTGCC...(unsequenced)....250bp..... 2057
Ala
250 260 270
alIleThrIleGlyAsnGluArgPheArgCysProGluAlaLeuPheGlnProSerPheLeuGlyMetGluSerCysGlyIleHisGluThrThrPheAs
cDNA : TGATCACCATTGGCAATGAGCGGTTCCGGTGTCCGGAGGCACTCTTCCAGCCTTCCTTCCTGGGCATGGAGTCCTGTGGTATCCATGAGACCACTTTCAA 700
λmA36: ...(unsequenced)...TATCCGGAGACACTCTTCAATCCTTCCTTCCTGGGCACGGATTCCTGTGGTATCCATGAGACCACCTTCAA 2371
Tyr Thr Asn ThrAsp
280 290 300
nSerIleMetLysCysAspValAspIleArgLysAspLeuTyrAlaAsnThrValLeuSerGlyGlyThr
cDNA : CTCCATCATGAAGTGTGATGTGGATATCCGCAAAGACCTGTATGCCAATACAGTGCTGTCTGGTGGTACC 770
λmA36: CTCCATCATGAAGTGTGATGTGGATATCCGCAAGGACCGGTATGCCAATACGGTGCTGTCTGGTGGTACC 2441
Arg

these regions there are 12 residues unique to cytoplasmic actin, and of these 11 were found to correspond to the predicted translation of the actin-like nucleotide sequence in λmA36. Only the amino-acid predicted at position 259 did not correspond to the non-muscle isotype. Therefore λmA36 most closely resembles a gene corresponding to a cytoplasmic isoform of actin (Vanderckhove and Weber, 1979a). However it cannot be determined from the region sequenced whether the cytoplasmic actin-like gene in λmA36 is of the β or γ isotype, as the 4 amino acids which identify the isotype are at the N-terminal end, and this region was not sequenced.

The actin-like gene of λmA14 bears some of the hallmarks of a processed pseudogene. There are 28 differences in the predicted amino acid sequence from that of γ-actin (represented by the residues below the λmA14 nucleotide sequence in Figures 4.1 and 4.2), including an Ile residue at position 1 rather than an initiating methionine, and stop codons rather than Arg and Tyr at positions 38 and 166, respectively. These changes clearly preclude this actin-like DNA from having any functional potential and identify it as a pseudogene.

The actin-like sequence in λmA14 (Figure 4.1 and 4.2) is not interrupted by the introns anticipated for mouse γ-actin. Although it is not known whether the gene coding for the mouse (or indeed any mammalian) γ-actin has introns, the genes for the four mammalian actin isoforms so far characterised all have introns at amino-acid positions 41, 267 and 327, as well as at other positions specific for different isoforms, (Carroll *et al.*, 1986; Chang *et al.*, 1984, 1985; Hamada *et al.*, 1982; Ng *et al.*, 1985; Bergsma *et al.*, 1985; Foran *et al.*, 1985; Table 1.2). Thus it seems most likely that mouse γ-actin will also possess introns and that the pseudogene in λmA14 is

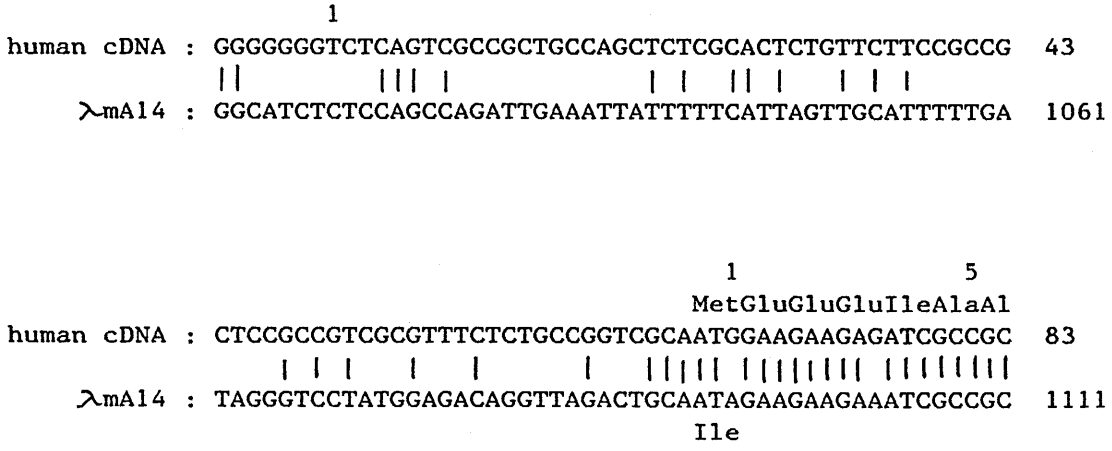
therefore of the processed type.

Most processed pseudogenes contain DNA copies of the whole of the mRNA, including the 5' and 3' untranslated regions, and a 3' poly A tract, and are flanked at the 5' and 3' ends by a short target-site direct repeat. The actin-like coding amino acid sequence in λmA14 probably includes the residue at position 1, as ATA may well be a mutated (ATG) initiation codon (Figure 4.1). The 5' untranslated region of the mouse cytoplasmic γ-actin gene has not yet been sequenced and therefore it is difficult to determine whether the actin pseudogene in λmA14 includes all or part of this region. However the nucleotide sequence of a 'full-length' human cytoplasmic γ-actin cDNA, including 73 bases of the 5' untranslated region is available (Erba *et al.*, 1986). As the 5' untranslated regions of the rat (Nudel *et al.*, 1983) and human (Ponte *et al.*, 1984) β-actin genes show more than 80% identity, it may therefore be valid to compare the λmA14 sequence with the 5' untranslated region of human γ-actin. Examination showed that the homology between these sequences appeared to extend only 3 bases to the left of the presumed actin-coding residue 1 (Figure 4.4). If this is the left-hand end of the actin pseudogene in λmA14, it would appear to be truncated, as one would expect the 5' untranslated region to be longer than 3 bases. A human cytoplasmic γ-actin processed pseudogene has recently been sequenced and it does not appear to be truncated at the 5' end (Leube & Gallwitz, 1986). However, an unequivocal 5' truncation of a mouse cytoplasmic γ-actin pseudogene has previously been described (Leader *et al.*, 1985), the actin-like sequence beginning at amino acid position 7.

The few previously reported examples of processed pseudogenes truncated at the 5' end, fall into several different categories. In some cases

Figure 4.4 Comparison of the 5' untranslated region of human γ-actin cDNA with the corresponding region in λmA14

The 5' untranslated nucleotide sequence of the human cytoplasmic γ-actin cDNA, pHFγA-1 (Erba *et al.*, 1986), is compared with the 5' flanking sequence of the γ-actin processed pseudogene in λmA14. The nucleotide sequence of pHFγA-1 is numbered as in Erba *et al.*, (1986) and the λmA14 nucleotide sequence is numbered as in Figure 3.22.


1
human cDNA : GGGGGGGTCTCAGTCGCCGCTGCCAGCTCTCGCACTCTGTTCTTCCGCCG 43
λmA14 : GGCATCTCTCCAGCCAGATTGAAATTATTTTTCATTAGTTGCATTTTTGA 1061
1 5
MetGluGluGluIleAlaAl
human cDNA : CTCCGCCGTCGCGTTTCTCTGCCGGTCGCAATGGAAGAAGAGATCGCCGC 83
λmA14 : TAGGGTCCTATGGAGACAGGTTAGACTGCAATAGAAGAAGAAATCGCCGC 1111
Ile

the 5' truncation is clearly explained by the insertion of another retroposon (Shimada *et al.*, 1984; Scarpulla *et al.*, 1984). In other cases the processed pseudogene appears to be derived from an abberrant transcript generated by faulty splicing or by initiation down-stream from the normal cap site. For example, the pseudogenes derived from the human immunoglobulin lambda light chain (Hollis *et al.*, 1982), and the human immunoglobulin epsilon heavy chain (Battey *et al.*, 1982; Ueda *et al.*, 1982). These examples are of genes that are subjected to strict tissue-specific regulation in the soma and may perhaps only give rise to pseudogenes from abberrant germline transcripts.. However there are two examples of genes which are not tissue-specific and give rise to 5' truncated pseudogenes, not caused by retrotransposon insertion. These are the mouse γ-actin pseudogene in λmA19 (Leader *et al.*, 1985) and a mouse cellular tumour antigen p53 pseudogene, where at least 80 nucleotides are missing from a long 5' untranslated region (Zakut-Houri *et al.*, 1983). It has been suggested that such genes may have arisen from incomplete or partially degraded reverse transcripts of a full-length mRNA. The 5' flanking nucleotide sequence of the actin processed pseudogene in λmA14, was compared with sequences in the EMBL databank, and was found not to be related to any of the entries, of retroposon origin or otherwise. Therefore if this gene is really truncated at the 5' end, it may have also arisen from an incomplete or degraded reverse transcript. If this is the case, the occurrence of two truncated mouse actin γ-actin processed pseudogenes may indicate that there is a large amount of secondary structure at the 5' end of the mouse γ-actin mRNA which is a barrier to reverse transcription *in vivo*.

Although it is not necessarily evident from direct comparison, the results of section 3.1 have already demonstrated that the actin-like genes in

λmA14 and λmA36 are parts of a much larger area of similar DNA in these clones, of at least 11.0kb in length. The question of the origin of this similarity (duplication or amplification) is discussed in 4.3, below. However it is convenient at this juncture to discuss the actin-like regions in λmA14 and λmA36 from the stand point of this relatedness.

It was of interest to determine whether the actin pseudogene regions of λmA14 and λmA36 showed similar divergence to that found throughout the rest of the duplicated/amplified DNA. The degree of similarity was determined by sequencing the available subclones at the leftward and rightward extremities of the similarity (Figure 3.20). Comparison of these sequences indicated that λmA14 and λmA36 have diverged by 4%. Comparison of the actin-like coding regions and the 5' flanking DNA (Figure 3.25 (a) and (b)), indicated that λmA14 and λmA36 have diverged by 6% in the actin-like region and 7% (which includes the leftward extremity sequence) in the 5' flanking region. The small differences observed for the percentage divergence is most likely due to DNA sequencing errors and/or comparison of relatively short lengths of sequence. Therefore it would be unwise to conclude from the data that there is any significant difference in the relatedness of λmA14 and λmA36 for the actin regions and for the flanking DNA.

We now turn to consider the evolutionary time-scale of the events which led to the formation of λmA14 and λmA36, by comparing the percentage divergence of the nucleotide sequence of the two clones and by comparing their γ-actin region with the nucleotide sequence of γ-actin cDNA nucleotide sequence.

To determine how long ago in evolutionary time λmA14 and λmA36

diverged, the average percentage sequence divergence (5.7%) was used, and the assumption was made that the DNA in these regions (pseudogenes and unknown flanking DNA) has evolved at a neutral rate, free from any selective pressure. The average neutral rate at which nucleotide substitution occurs in pseudogenes, was shown by Li *et al.*, (1981), to correspond to a UEP (unit evolutionary period) of 0.46, or a mutation rate of 4.6 X 10^{-9} substitutions per nucleotide per year. (UEP is the time in millions of years (MY) required for the fixation of 1% changes between two lines; Perler *et al.*, 1980). Therefore assuming neutral drift for the sequences compared in λmA14 and λmA36, it can be concluded that these DNAs diverged approximately 2.6 MY ago (5.7% with a UEP of 0.46), presumably as the result of a gene duplication or amplification event at that time. It should be stressed that the validity of this conclusion is limited by the assumption of neutral drift at the rate found in α-globin pseudogenes (Li *et al.*, 1981), and the accuracy with which the figure 5.7% represents the true divergence of λmA14 and λmA36.

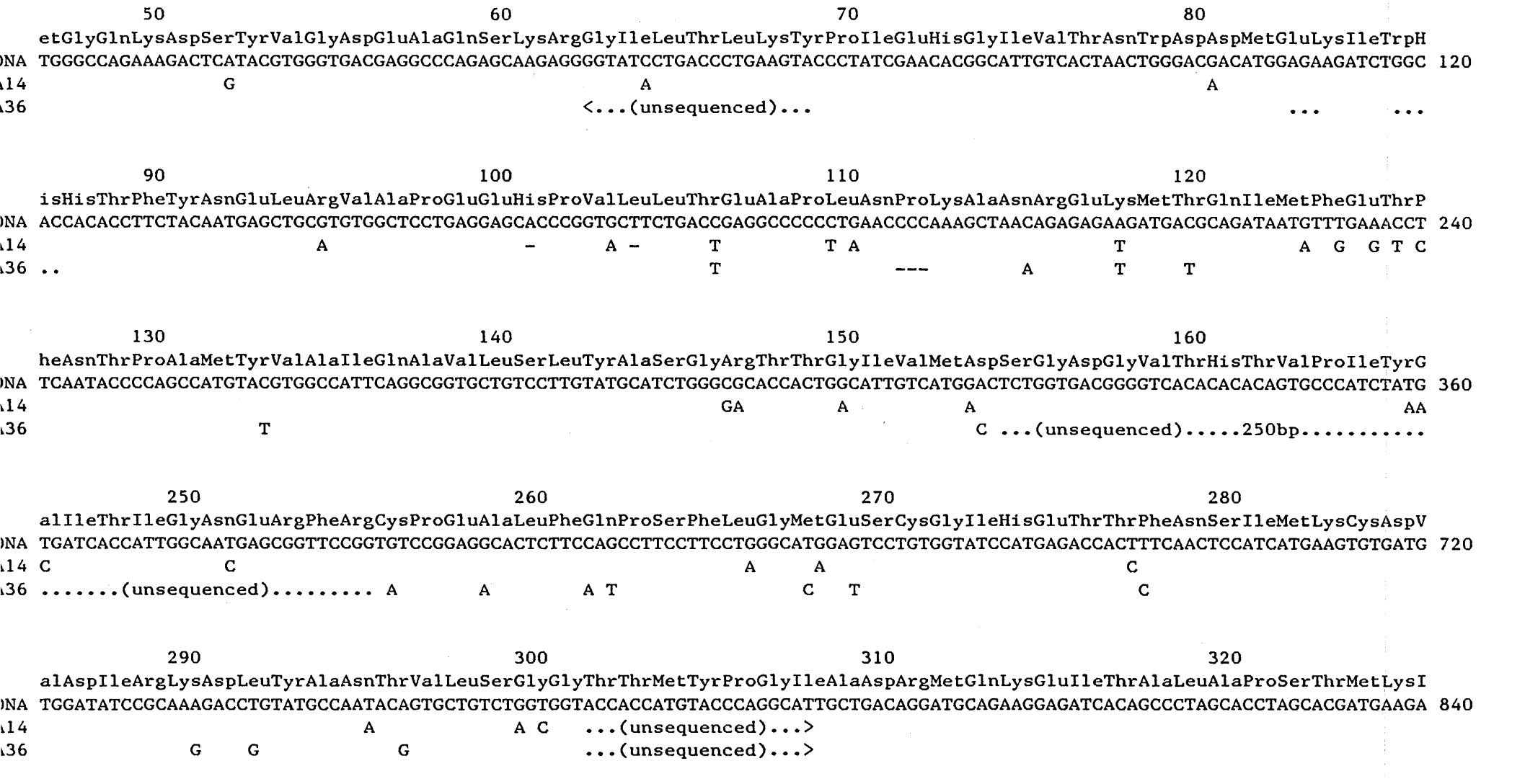
To determine how long ago in evolutionary time the γ-actin DNA in λmA14 and λmA36 diverged from the active γ-actin gene the percentage divergence of these sequences from that of a recently sequenced γ-actin cDNA (Peter and Leader, unpublished), was calculated. Changes in the nucleotide sequence could only be determined in the region corresponding to amino acids 48 onward, for which both pseudogene and cDNA sequence were available. Each base change was scored as 1, as was each insertion/deletion, irrespective of size. Comparison of the nucleotide sequences showed that the actin-like gene in λmA14 was 94.9% identical to the cDNA sequence (Figure 4.2). Comparison of the partial nucleotide

sequence of the actin-like gene in λmA36 with the cDNA sequence from amino acid residues 70 to 157 and 256 to 302, indicated that the sequences were 95.8% identical (Figure 4.3). The average percentage divergence of λmA14 and λmA36 γ-actin DNA from the γ-actin cDNA sequence (4.65%) was used to calculate the time of divergence using the assumption that the actin pseudogenes had evolved at a neutral rate since their formation and the gene (as represented by the cDNA sequence) had evolved under selective pressure for a protein coding sequence. This gave a UEP value for the divergence of the gene and pseudogene of 0.81, from which it is concluded that the γ–actin genes in λmA14 and λmA36 diverged from the active γ-actin gene approximately 3.8 MY ago (4.65% with a UEP of 0.81). As the duplication or amplification event was calculated above to occur approximately 2.6 MY ago, this suggests that the original actin processed pseudogene represented in λmA14 or λmA36 existed for approximately 1.2 MY before it was duplicated/amplified. However if this were the case one would expect that mutations aquired over this postulated first 1.2 MY would be common to λmA14 and λmA36 and would represent approximately 30% of the total. In fact only 2 out of 39 differences from the γ-actin cDNA sequence are common to both λmA14 and λmA36 (Figure 4.5), more in accord with a duplication / amplification event occurring much sooner after the original pseudogene emerged. The cause of this discrepancy is unclear.

To determine whether the actin-like sequences in λmA14 and λmA36 have evolved at a neutral rate, as assumed for the calculations above, the R/S ratios were calculated. In a functional coding sequence, R (replacement) changes are more likely to be det rimental and therefore selected against rather than S (silent) changes. As a consequence, the R/S ratio, can be used

Figure 4.5 Comparison of mutations in the actin pseudo-coding region of λmA14 and λmA36

The diagram shows a comparison of, the base mutations which have occurred in the actin pseudo-coding region of λmA14 and λmA36 (residues Pro^{70} to Asp^{157} and Tyr^{256} to Thr^{302}) and the γ-actin cDNA nucleotide sequence (Peter & Leader, unpublished). Unsequenced regions are indicated by dots (••••••) and nucleotide base deletions are indicated by a dash (-).


50
60
70
80
etGlyGlnLysAspSerTyrValGlyAspGluAlaGlnSerLysArgGlyIleLeuThrLeuLysTyrProIleGluHisGlyIleValThrAsnTrpAspAspMetGluLysIleTrpH
ONA TGGGCCAGAAAGACTCATACGTGGGTGACGAGGCCCAGAGCAAGAGGGGTATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGACGACATGGAGAAGATCTGGC 120
A14 G A A
A36 <...(unsequenced)... ... ...
90
100
110
120
isHisThrPheTyrAsnGluLeuArgValAlaProGluGluHisProValLeuLeuThrGluAlaProLeuAsnProLysAlaAsnArgGluLysMetThrGlnIleMetPheGluThrP
ONA ACCACACCTTCTACAATGAGCTGCGTGTGGCTCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAAGCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCT 240
A14 A - A - T T A T A G G T C
A36 .. T --- A T T
130
140
150
160
heAsnThrProAlaMetTyrValAlaIleGlnAlaValLeuSerLeuTyrAlaSerGlyArgThrThrGlyIleValMetAspSerGlyAspGlyValThrHisThrValProIleTyrG
ONA TCAATACCCCAGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGCGCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTGCCCATCTATG 360
A14 GA A A AA
A36 T C ...(unsequenced).....250bp.....
250
260
270
280
alIleThrIleGlyAsnGluArgPheArgCysProGluAlaLeuPheGlnProSerPheLeuGlyMetGluSerCysGlyIleHisGluThrThrPheAsnSerIleMetLysCysAspV
ONA TGATCACCATTGGCAATGAGCGGTTCCGGTGTCCGGAGGCACTCTTCCAGCCTTCCTTCCTGGGCATGGAGTCCTGTGGTATCCATGAGACCACTTTCAACTCCATCATGAAGTGTGATG 720
A14 C C A A C
A36 ....(unsequenced)..... A A A T C T C
290
300
310
320
alAspIleArgLysAspLeuTyrAlaAsnThrValLeuSerGlyGlyThrThrMetTyrProGlyIleAlaAspArgMetGlnLysGluIleThrAlaLeuAlaProSerThrMetLysI
ONA TGGATATCCGCAAAGACCTGTATGCCAATACAGTGCTGTCTGGTGGTACCACCATGTACCCAGGCATTGCTGACAGGATGCAGAAGGAGATCACAGCCCTAGCACCTAGCACGATGAAGA 840
A14 A A C ...(unsequenced)...>
A36 G G G ...(unsequenced)...>

to discriminate between functional genes and pseudogenes, pseudogenes being in general expected to have 2.5-3.0 times as many R as S changes because there are more potential sites for R changes (Czelusniak *et al.*, 1982). In the case of mammalian actins, with their absolutely conserved amino acid sequences, one can calculate the expected R/S ratio for a pseudogene precisely. In the γ-actin pseudogenes under consideration here one would expect to have 3.0 times as many R changes as S changes. Analysis of the λmA14 actin-like sequence indicates that there are 39 base diferences from that of the gene, 15 at silent sites and 24 at replacement sites, producing a R/S ratio of 1.6. The actin-like sequence in λmA36 had 17 base changes, 5 at silent sites and 12 at replacement sites, producing a ratio of 2.4. Given the small total number of base changes, one can say that the actin-like gene in λmA36 had approximately the predicted ratio for a pseudogene evolving under neutral selection. However the R/S ratio for the actin-like gene in λmA14 is intermediate between the value expected of a functional gene and that of a pseudogene. Taken at face value, this would suggest that during part of its existence, the gene has evolved under selective pressure to conserve a protein-coding potential. However this seems unlikely as most processed pseudogenes, being derived from transcripts lacking a RNA polymerase II promoter, are expected to be inactive as soon as they arise. There is an example of an 'active' calmodulin processed pseudogene (Stein *et al.*, 1983), apparently fortuitously inserted after a polymerase II promoter. Inspection of the 5' sequence flanking the coding region of λmA14 reveals no such promoter, although it cannot be excluded that one existed for a time and was subsequently deleted. Another possible explanation for this anomalous R/S value could be that the actin-like sequence has been subjected to gene conversion. However, if so

this could not have involved a γ-actin sequence as the actin sequence in λmA14 is as diverged from the γ-actin cDNA sequence as the γ-actin sequence in λmA36. Thus the reason for the low R/S ratio of the actin pseudogene in λmA14 remains unclear.

4.2 L1Md sequence in λmA14 and λmA36

The stimulus for the work described in this thesis was the observation of large foldback structures associated with the actin-like genes in λmA14 and λmA36 and the possibility that they represented discrete functional elements. It has been shown that these foldback structures are actually composed of L1Md sequences. Nevertheless it is pertinent to discuss the structure of the L1Md sequences, in relation to the possibility that they may be in some way related. In itself the proximity of the λmA14 LINEs does not necessarily suggest a relationship between them. For example, there are at least eight LINE members in the mouse β-globin region, each having a different length and being flanked by different direct repeats, suggesting that they inserted as separate elements (Voliva *et al.*, 1984; Shyman *et al.*, 1985). Also the evidence is conclusive that the three LINE members did not insert into the λmA14 DNA as a mobile unit. L1Md-LH had inserted into the actin pseudogene of λmA14 displacing its' extreme 3' end at least 3.3kb to the right, indicating its insertion must have been an independent event. However, as discussed in detail below, the LINE members in λmA14, particularly L1Md-LH and L1Md-RH1, do in fact share some common sequence characteristics which suggests a relationship between them.

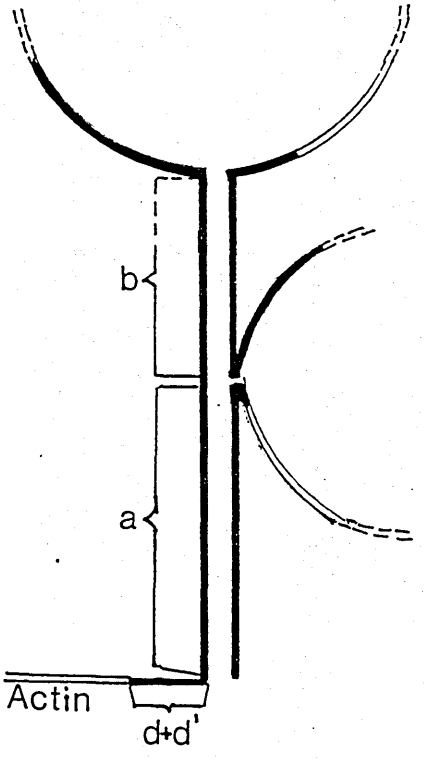
Before examining these common characteristics, it may be useful to consider how the electron micrograph foldback structures can be accounted for by the structural data described in detail in section 3.2. Figure 4.6 shows the stem regions of the foldback structures within λmA14 and λmA36, divided into sections designated a to e, each shown with the original measurements predicted from the electron micrographs. The stem of the foldback structure of λmA14 is composed of specific regions of the three λmA14 L1Md members shown in Figure 4.7. (L1Md-LH on the left-hand side and L1Md-RH1 and L1Md-RH2 on the right-hand side of the stem). The original electron micrograph measurements predicted for each stem section agree reasonably well with those obtained by sequencing (Figure 4.7), although in each case the sequencing measurements are greater than those predicted by electron microscopy. This is to be expected if the full potential for hybridisation is not realised in practise. The precise length of section b is still unknown as the 3' end of L1Md-RH2 remains to be located. However partial sequencing of the 5' end of L1Md-RH2 appears to indicate that the first 460bp at this region are homologous to the 3' end of L1Md-RH1. The electron micrograph of λmA14 (Figure 1.5) may be interpreted in the terms of the 460bp of overlapping sequence in L1Md-RH2 being unable to hybridise to L1Md-LH as the complementary region in L1Md-LH had already hybridised to L1Md-RH1 forming stem section a. Thus only the 3' end of L1Md-RH2 hybridised to L1Md-LH, consitituting stem section b.

As the foldback structure in λmA14 is located within a large region of similar DNA to that represented in λmA36, the foldback structure in λmA36 can be discussed in relation to that in λmA14, even though no sequence determination was performed on the LINE members in λmA36. The electron

Figure 4.6 Diagrammatic representation of the λmA14 and λmA36 foldback stem regions

The stem regions of the foldback structures within λmA14 and λmA36, are designated into sections a to e. The position of the L1Md DNA, as determined by sequencing, is shown as solid black lines.

λmA14



λmA36

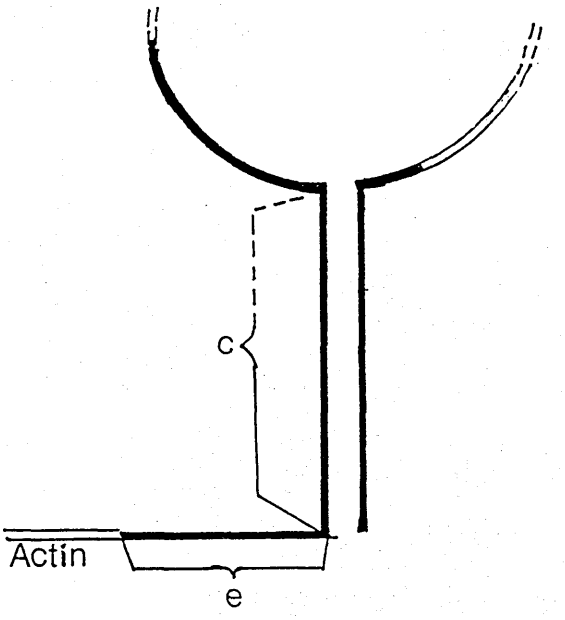
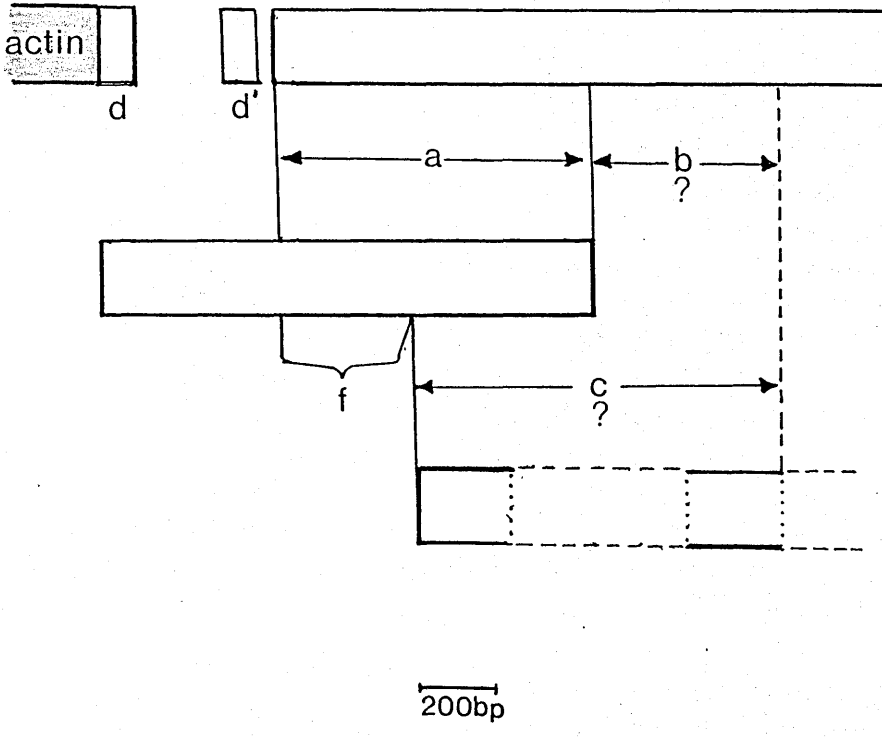


Figure 4.7 Linear representation of the λmA14 and λmA36 foldback stem regions

The stem regions of the foldback structures within λmA14 and λmA36 are designated into sections a to e, as described in Figure 4.6. The length of each section, predicted from the electron micrographs, is compared with the lengths determined by DNA sequencing. The stem of the foldback structure in λmA14 is composed of specific regions of LM1d-LH, L1Md-RH1 and L1Md-RH2. The stem of the foldback structure in λmA36 is composed of specific regions of L1Md-LH and L1Md-RH2. The λmA14 L1Md members are aligned so as to indicate the regions of homology.



LIMd-LH

LIMd-RHI

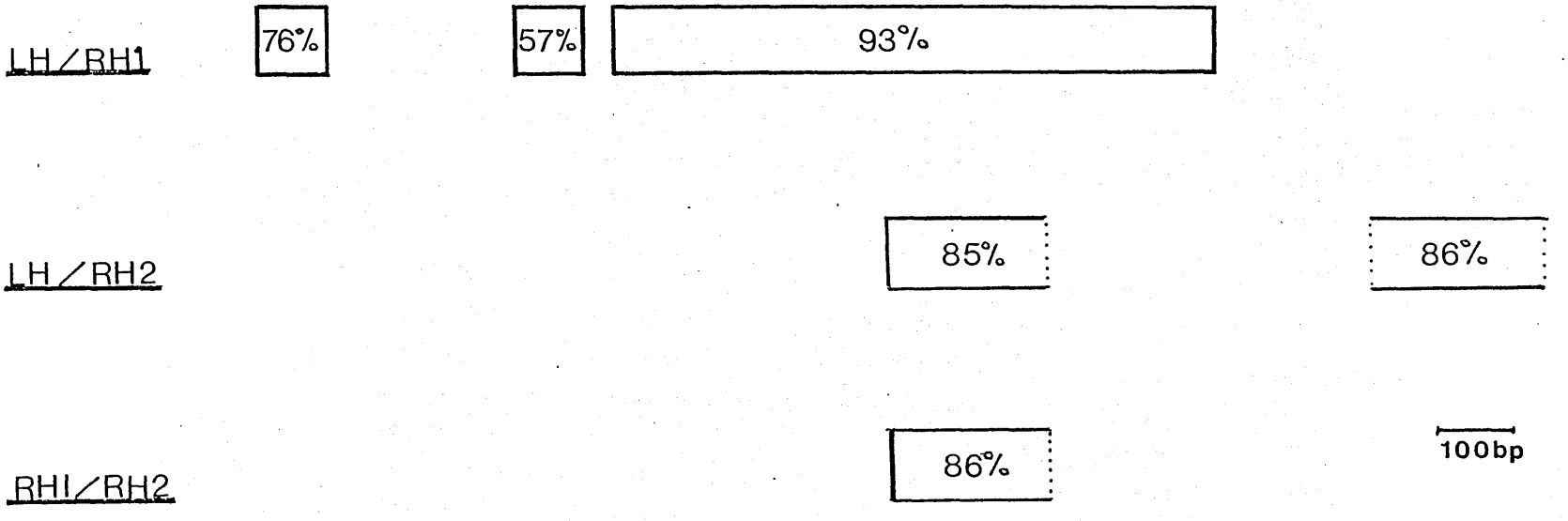
LIMd-RH2

Region	E.M.	DNA Seq.
a	760	820
b	540	480 ?
c	870	940 ?
d+d'	50	190
e $\left(\begin{smallmatrix} d+d' \\ f \end{smallmatrix}\right)$	550	500

micrographs show that although λmA14 and λmA36 contain a similar foldback structure they differ in detail. The smaller clone λmA36 does not contain the more extreme right-hand L1Md member, L1Md-RH1', which if present in the genomic region represented by λmA36 would lie outwith this clone. The stem of the foldback in λmA36 was thus composed of only L1Md-LH' and L1Md-RH2'. In the absence of L1Md-RH1', all of L1Md-RH2' would be expected to hybridise to L1Md-LH', and this partly accounts for the way that the structure of the foldback in λmA36 differs from that in λmA14. The foldback stem in λmA36, was displaced further to the right of the actin region (section e), at a distance predicted from the electron micrograph to be 550bp and calculated to be 500bp by DNA sequencing. The foldback structure in λmA36 did not contain a side loop and its stem length (section c) was estimated to be 870bp. As the 3' end of L1Md-RH2' has not yet been located, the precise length of the stem section c remains undetermined. However at this stage L1Md-RH2' is predicted to contain at least 940bp of DNA complementary to L1Md-LH'.

It is necessary now to turn to a detailed comparison of the structures of the three L1Md members, of λmA14 to discover whether they are related in any way. The nucleotide sequences of the three λmA14 L1Md members are compared in Figures 3.38 to 3.40. The percentage homology between the LINE members was calculated for the different regions and is shown diagrammatically in Figure 4.8. The division into different regions was made because L1Md-LH has a deletion near its 5' end, which is not total. There is a small block remaining which, although it did not match very easily to L1Md-RH1 (or as shown later L1Md-A2) was assigned for convenience to an area with which it shared approximately 55% identity. Each base change is scored

Figure 4.8 Diagrammatic representation of the percentage homology between the λmA14 L1Md members



as 1, as is each insertion/deletion irrespective of size. The greatest region of homology was found between L1Md-LH and L1Md-RH1 and was 93%, however towards the 5' end at the 5' side of the 266bp deletion in L1Md-LH at position 99 (Figure 3.38), the homology was reduced to 76% (As would be anticipated from the above, the small remnant in the deletion of L1Md-LH had only low (57%) apparent homology to L1Md-RH1). The homology between L1Md-LH and L1Md-RH2 (Figure 3.39), L1Md-RH1 and L1Md-RH2 (Figure 3.40), was similar and approximately 86%, (Figure 4.8).

To put the figures above in perspective, comparison of these λmA14 LINE members was made with a 'full-length' L1Md member designated L1Md-A2 (Loeb *et al.*, 1986), as shown diagrammatically in Figure 4.9. It transpires that overall each λmA14 L1Md member is slightly more homologous to L1Md-A2, than it is to the others. Although interpretation of these figures is not easy, they do not immediately suggest a relationship between the λmA14 L1Md members. However a more detailed examination of the nucleotide sequences of L1Md-LH and L1Md-RH1, does reveal that they possessed several features in common. One is that they share the same 5' ends, (Figure 4.10). This point is of some significance because although assumed 'full-length' L1Md members have tandem 208bp repeats at their 5' ends, the number of these varies. The examples so far described have 4 $^{2}/_{3}$ (L1Md-A2) and 1$^{2}/_{3}$ copies (L1Md-9), the $^{2}/_{3}$ copy being the most 5' member and the exact position at which the $^{2}/_{3}$ copy starts being slightly different (Loeb *et al.*, 1986). L1Md-LH and L1Md-RH1 have approximately 1 $^{2}/_{3}$ copies of the 5' tandem repeats and the exact position at which the $^{2}/_{3}$ copy starts in these two examples appears to be exactly the same (Figure 4.10) and differs from both L1Md-A2 and L1Md-9 which are 16 and 6 nucleotides longer, respectively. Although the 5' end point of L1Md-LH is obvious from

Figure 4.9 Diagrammatic representation of the percentage homology between the λmA14 L1Md members and L1Md-A2

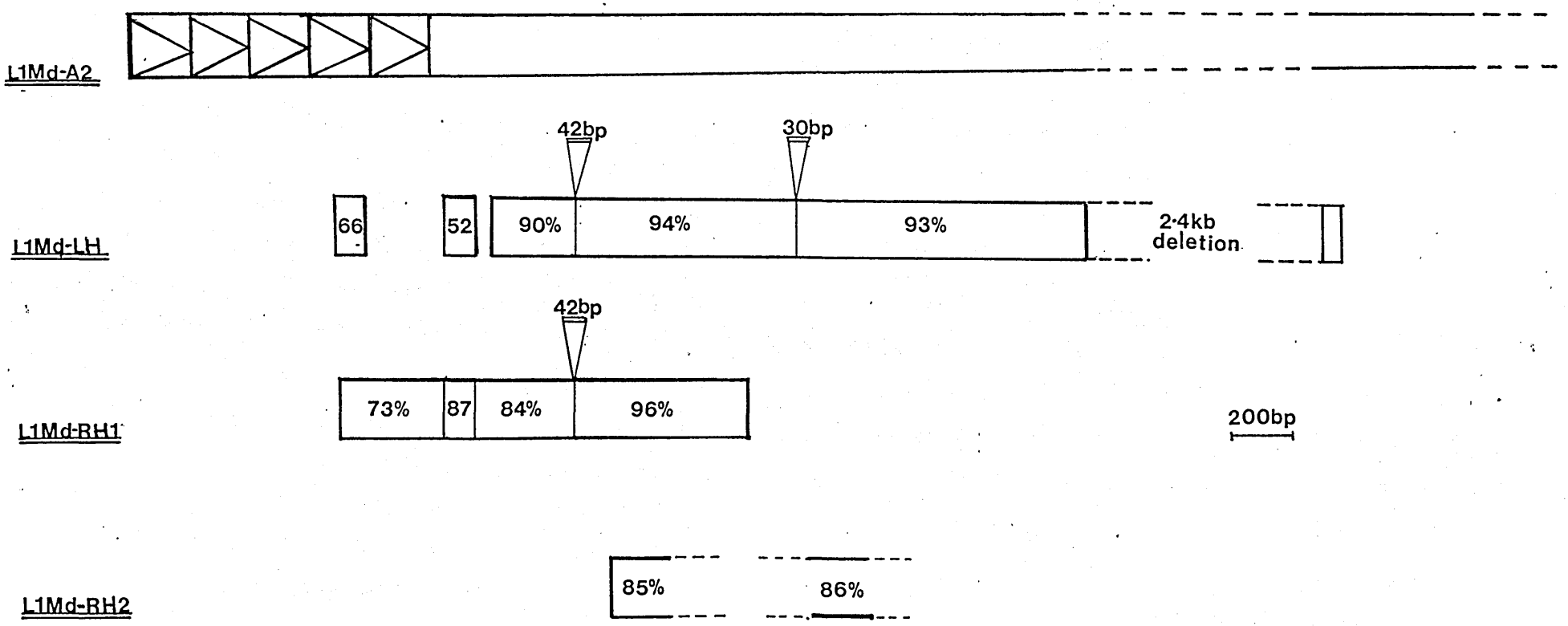
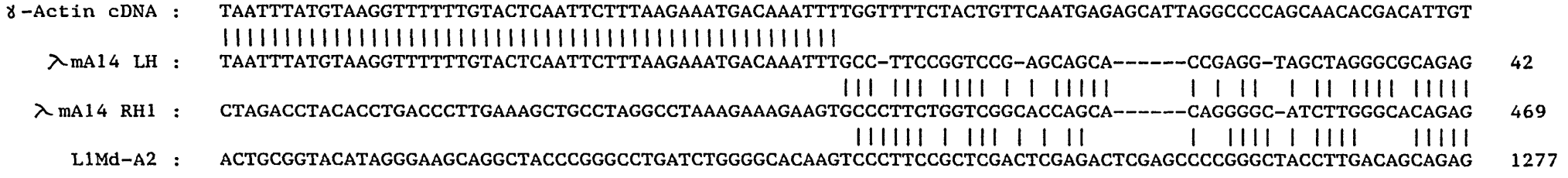


Figure 4.10 Comparison of the 5' ends of L1Md-LH and L1Md-RH1

The nucleotide sequence at the 5' ends of L1Md-LH and L1Md-RH1 is compared with the 3' non-coding γ-actin cDNA (Peter & Leader, unpublished) and L1Md-A2 (Leob *et al.*, 1986) nucleotide sequence. L1Md-LH and L1Md-RH1 sequence are numbered as in Figure 3.33 and 3.35. The L1Md-A2 sequence is numbered in accordance with Loeb *et al.*, (1986).


ϑ-Actin cDNA : TAATTTATGTAAGGTTTTTTGTACTCAATTCTTTAAGAAATGACAAATTTTGGTTTTCTACTGTTCAATGAGAGCATTAGGCCCCAGCAACACGACATTGT
λmA14 LH : TAATTTATGTAAGGTTTTTTGTACTCAATTCTTTAAGAAATGACAAATTTGCC-TTCCGGTCCG-AGCAGCA------CCGAGG-TAGCTAGGGCGCAGAG 42
λmA14 RH1 : CTAGACCTACACCTGACCCTTGAAAGCTGCCTAGGCCTAAAGAAAGAAGTGCCCTTCTGGTCGGCACCAGCA------CAGGGGC-ATCTTGGGCACAGAG 469
L1Md-A2 : ACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTCGACTCGAGACTCGAGCCCCGGGCTACCTTGACAGCAGAG 1277

the interruption of the actin-like sequence, it might appear that there is some ambiguity in that of L1Md-RH1, as insertion of an extra base in L1Md-A2 could produce a further four nucleotides at its 5' end. The argument against this is that the region 5' to the position designated as the 5' end of L1Md-RH1 is part of a repeating sequence. A diagonal matrix comparison plot of this 5' flanking DNA of L1Md-RH1 illustrating this is shown in Figure 4.11, and four imperfect direct tandem repeats of approximately 150bp of non-L1Md DNA are shown in Figure 4.11, the first base to the left of the 5' end of L1Md-RH1 being part of a repeat unit. This 150bp repeat does not correspond to a different repeating sequence found at the 5' end of certain L1Md members (Fanning, 1983), nor to any other sequence in the EMBL and GenBank databanks.

The second point of similarity was also at the 5' end. Although comparison here was restricted to 105bp because of a deletion after this point, L1Md-LH and L1Md-RH1 both display a greater percentage identity to one another (76%) than to the L1Md-A2 sequence (66% and 73%, respectively) in this region. Also within this 5' region L1Md-LH and L1Md-RH1 share similar single base deletions relative to L1Md-A2 at three positions and a common 6bp deletion (Figure 3.43). This greater similarity of L1Md-LH and L1Md-RH1 to one another at the 5' end contrasts with the pattern over the rest of their lengths as described above.

A third way in which L1Md-LH and L1Md-RH1 differ from L1Md-A2 is in having an extra 42bp relative to L1Md-A2 at corresponding positions (519 and 1189, respectively). This presumed insert is an imperfect repeat of the preceding region (Figure 3.43), and therefore probably arose by tandem duplication. However, a deletion in L1Md-A2 cannot be excluded.

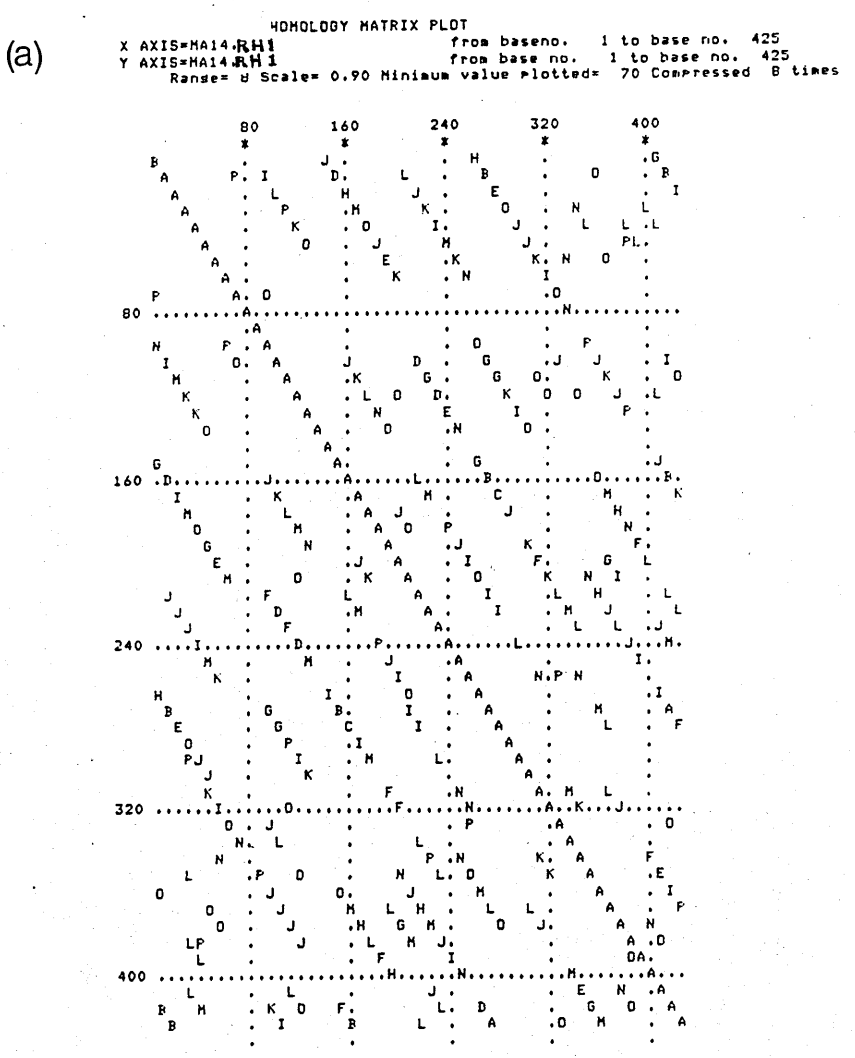
Taken together, these similarities are suggestive of a relationship between L1Md-LH1 and L1Md-RH1. It is possible that at some stage there may

Figure 4.11 Imperfect direct tandem repeats within the 5' flanking DNA of L1Md-RH1

(a) A diagonal matrix comparison plot of the 5' flanking DNA of L1Md-RH1 (nucleotides 1 to 425, as numbered in Figure 3.35).

(b) The four imperfect direct tandem repeats of non-L1Md DNA located within the 5' flanking DNA of L1Md-RH1. The repeats are numbered in accordance with Figure 3.35.

(a)



(b)

```
22    GATATGTGTCTAGGCCTATTCCTGAACATTGAAGAGGCCCGGACTAAAAGCAAAATAGTTGAGG--   85
      ||| || | ||||||| || ||||| | ||||||||||||     | |||||| | |||  | |
165   GATGTGGGCCTAGGCC-ATCCCTGACCCTTGAAGAGGCCCTACGCA-AAGCAAGA-AGTGAAAG-T  226
      ||||||||  | |||| ||||| | | ||||||||||||| | |  | ||||||||||| | |||| |
278   GATGTGGGTGTGGGCCTATCCCAGTCACTTGAAGAGGCCTTGGCAATAAGCAAGAAATTTAAAGAT  344


1                                                          CCCGGGCCTAAAGAAAGAAGA   21
                                                           ||  ||||||||||||||
86    GC-TAGGGTCAAAAGCAAGAAGTG-AGGGGGCCTAGGTCTATCCCAGATCTTTGTTTGAATCCTA-GCCTAAAGAAAGAATT  164
                              |||||||||||| || ||  ||||   ||| ||   |||||||||||||| |
227   ------------------------------GCCTAGGTCTATACCTGACTTTTGAA-GAAGCCAGAGCCTAAAGAAAGAAGT  277
                              ||||||  ||| |||||||  |||||  |  |||   ||||||||||||||||
345   GCCTAGGCCCAAT-GCAAGACGTGAAGAGGGCCTAGACCTACACCTGACCCTTGAAAGCTGCCTAGGCCTAAAGAAAGAAGT  425
```

have been a gene conversion event between L1Md-LH and L1Md-RH1. If so, this would have to have been long enough ago in evolutionary time to allow considerable subsequent divergence to occur.

L1Md-LH appears to be a 'full-length', if somewhat, internally mutated LINE member, and L1Md-RH1 seems to represent a LINE member with an intact 5' end. As most of the L1Md members described previously appear more or less severly truncated at their 5' ends, it is therefore worth examining these in more detail. As discussed above, Figure 4.9, shows each individual L1Md member in λmA14 compared with L1Md-A2 of Loeb *et al.*, (1986). The homology for all three LINE members is lowest at the 5' end, in the vicinity of the tandem repeats. This is not an unexpected observation as even within a single L1Md member these repeats are not identical, and differences were also observed between L1Md-A2 and another LINE member, L1Md-9, that was partially sequenced (Loeb *et al.*, 1986). All three λmA14 LINE members display the greatest homology within the region corresponding to the first of the two postulated protein-coding sequences of L1Md-A2 (the second of these is not sufficiently represented here to comment upon), consistent with this postulate.

Finally some other points about the LINE members in λmA14 require comment. One is that there are 14bp and 29bp insertions into positions 994 and 1289 (Figure 3.43), repectively of L1Md-LH. In the 14bp insertion, 11bp of the inserted DNA was a repeat of the following region. As with the 42bp insertion discussed above, this insertion is assumed to be the product of a tandem duplication. Internal duplication has been observed in other LINE members, for example 'R4' (Gebhard *et al.*, 1982, 1983). The 29bp insertion was unusual in having a sequence $A(CA_3)_7$, which is similar to that of a retroposon tail. In all primate and rodent retroposon classes, the 3' tails of

retroposons usually have the structure A_n or $(NA_x)_y$, where N is most often C, (Rodgers, 1985). It is proposed that this insertion is the remnants of a retroposon which has inserted into and then out off L1Md-LH.

As regards L1Md-RH1, it is necessary to account for the truncation at its 3' end. The obvious possibility is that an original full-length L1Md sequence suffered a massive deletion of its 3' portion. However an alternative mechanism has been suggested to explain the occurrence of other LINE fragments (usually deleted at both ends, as appears to be the case of L1Md-RH2). This is via non-homologous recombination involving L1Md sequences originating from extra chromosomal DNA circles which carry either LINE sequence alone (Schindler and Rush, 1985) or in association with short or long segments of non-LINE DNA circles (Jones and Potter, 1985; Fujimoto *et al.*, 1985). However there is no data to support or reject either proposition for the nature of the truncation of L1Md-RH1 and L1Md-RH2.

4.3 Amplification / duplication of λmA14 and λmA36

It is now proposed to address the question of how the two related genomic regions of the clones λmA14 and λmA36 arose. It would seem that there are three major alternatives :

(1) Two actin genes independently inserting at the same point into a similar stretches of DNA.

(2) A tandem duplication involving the generation of these two genomic regions from a single original genomic region which contained one of them.

(3) An amplification event in which a number of similar regions, including those represented in λmA14 and λmA36 were generated.

The first alternative seems least likely to be correct as processed pseudogenes examined to date do not have preferred DNA target sequence for insertion, although they do have a tendency to insert into A-rich regions (Rogers, 1985). The probability of (1) occurring by chance would be effectively zero.

Although it is not possible to distinguish definitively between the alternatives (2) and (3), certain facts bear upon them. The initial screening of the mouse genomic library, that yielded λmA14 and λmA36, also gave eight other clones containing actin processed pseudogenes, which were analysed by electron microscopy (by Dr H. Delius). However inverted repeat structures of the type seen in clones λmA14 and λmA36, were not observed in the others. Clones λmA14 and λmA36 were selected at moderate hybridisation stringency, conditions which have been shown to yield about 15 - 20 mouse actin pseudogenes, presumably representing both β and γ isoforms (Minty *et al.*, 1983). The occurrence of the foldback structure in only two out of the ten actin clones would thus make an extensive amplification appear unlikely. By elimination it therefore seems likely that the similar regions in λmA14 and λmA36 are most probably the result of a large tandem duplication. Chromosome walking is needed to determine whether this conclusion is correct. Other experiments using a non-L1Md, non-actin part of the common region of λmA14 and λmA36, as a probe would also be useful, as they would address the question of amplification.

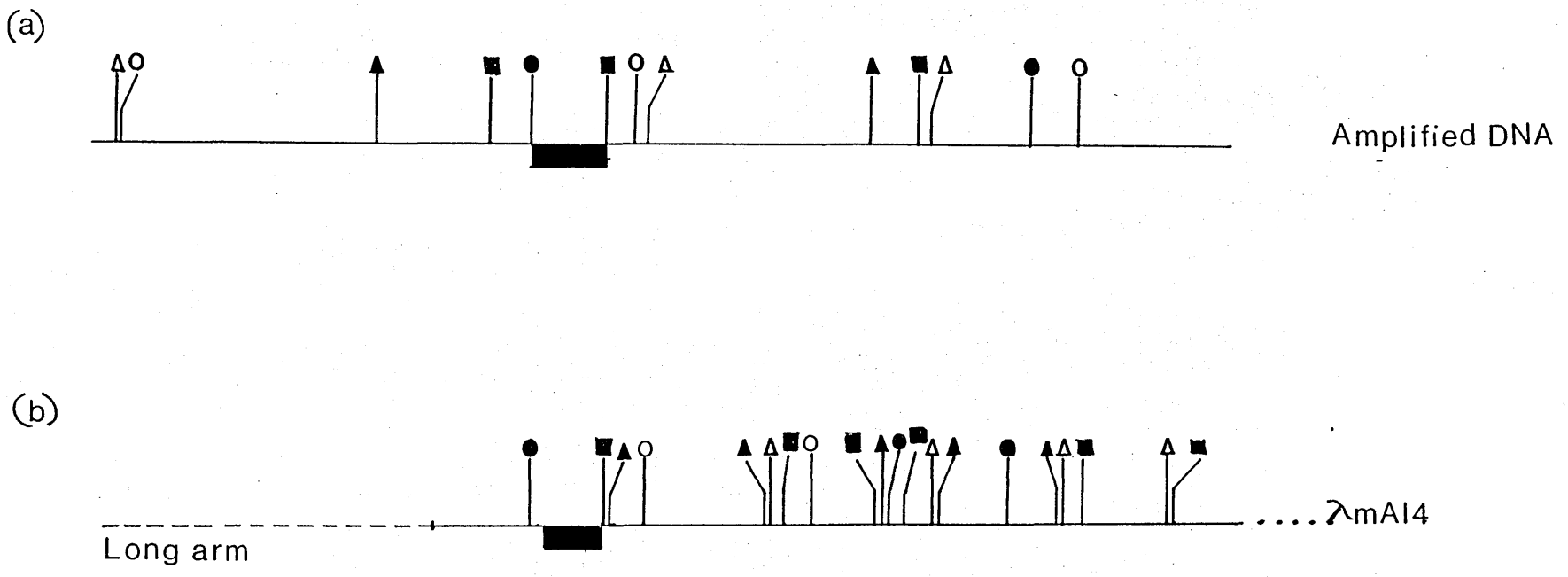
Despite this conclusion it must be pointed out that extensive amplification of the mouse genome involving actin-related sequences has

been described by Minty *et al.*, (1983). A sub-family was identified of sequences distantly related to a β-actin cDNA probe. The stringency at which these sequences were detected was such as to indicate that they were greater than 20% diverged from the actin sequence. This sub-family had resulted from the recent 20 - 50 fold amplification of a 17kb region of mouse genomic DNA. However it is clear that this amplified region does not correspond to the repeated regions in λmA14 and λmA36. This conclusion is based on two considerations. Firstly the actin-like DNA in λmA14 and λmA36 was detected at a higher stringency than could be used to detect the actin sequence of the sub-family. This indicated (and was subsequently confirmed by DNA sequencing) that the actin-like DNA within λmA14 and λmA36 is more closely related to the actin cDNA sequence (5%) than that of the actin-like sequence in the sub-family (> 20%). Secondly the predicted restriction map of the amplified sub-family DNA in the mouse genome differs extensively from that of the corresponding region of λmA14 (Figure 4.12). Nevertheless the studies described here are similar to those of Minty *et al.*, (1983), in that they provide evidence for recent evolutionary events involving mouse actin pseudogenes that partly account for the large number of actin-related sequences in this organism. Another amplification of an area of the mouse genome containing a processed pseudogene has been described: in this case involving a 45kb region containing a major urinary protein and its pseudogene (Clark *et al.*, 1985; Ghazal *et al.*, 1985).

The precise length of the DNA duplicated has not been determined because of the limited size of the cloned mouse DNA, in λmA14 and λmA36 and their incomplete analysis. However comparison of the restriction maps of λmA14 and λmA36 (Figure 3.15), indicates a loss of similarity occurring

Figure 4.12 Comparison of the restriction map of the mouse amplified region with that of λmA14

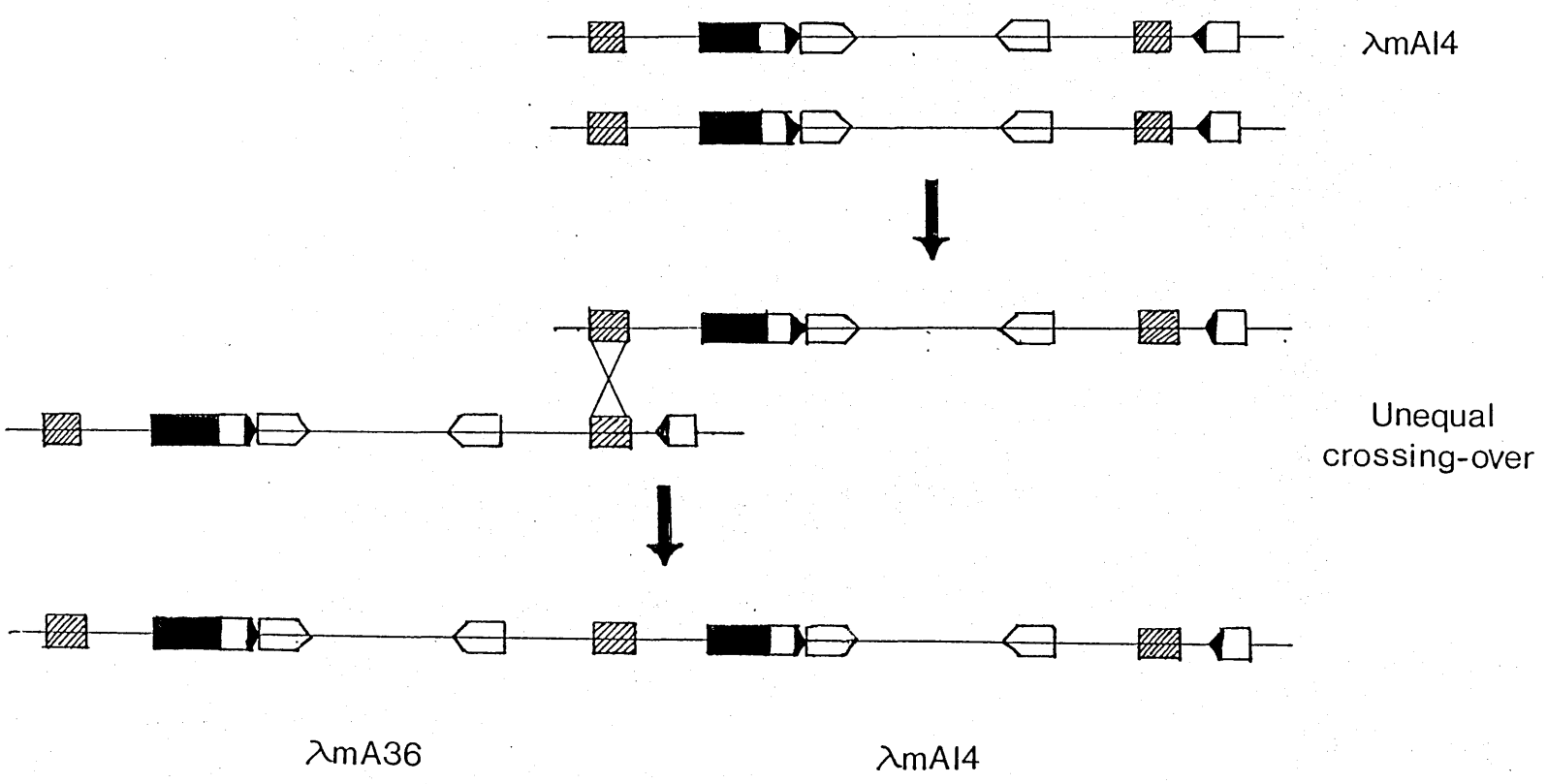
(a) The consensus restriction map of the amplified sub-family of actin sequences in the mouse genome, (Minty *et al.*, 1983) is compared with (b) the corresponding region in λmA14. Only the restriction sites for EcoRI (●), SstI (o), HindIII (▲), BamHI (▵) and XbaI (■) are shown. The solid area represents the position of the actin pseudo-coding region.


(a)
Amplified DNA
(b)
λmAI4
Long arm

within 1.5kb from the extremity of the short arm of the vector in λmA36. Loss of similarity in this region had also been inferred from electron microscopy which detected a 700bp non-looped inverted repeat in λmA36 (Figure 1.9), but not in λmA14. Whether this represents the vicinity of the true rightward end of the similarity or simply an interruption, is unknown. At the leftward end the similarity between λmA14 and λmA36 extends at least up to 1.2 and 0.7kb resepectively, from the left-hand lambda arm, at the position of the most leftward restriction sites mapped (Figure 3.15). Therefore all that can be concluded at the moment is that λmA14 and λmA36 share at least 11.0kb of similar DNA, which is most likely the product of a gene duplication event. Gene duplication events have been most clearly characterised in the human globin gene family. The lengths of DNA which have been duplicated in the formation of this family vary considerably. In the case of the human $^{G}\gamma/^{A}\gamma$, δ/β, $\psi\ \zeta_1/\zeta_2$ and α_1/α_2 pairs, duplicated regions of approximately 5, 7, 12 and 4kb DNA are still evident (Proudfoot *et al.*, 1982; Lauer *et al.*, 1980; Shen *et al.*, 1981; Martin *et al.*, 1983). Chromosomal walking is needed to determine the total length of the similar DNA associated with λmA14 and λmA36, and this would also establish whether these regions are in tandem.

Finally it may be asked whether the L1Md members within λmA14 and λmA36 could have been involved in the duplication event assumed to have given rise to these clones. The presence of these highly repeated DNA sequences within the duplicated unit could provide excellent targets for unequal crossing-over. Figure 4.13, shows an example of how unequal crossing-over between L1Md members 5' and 3' to the actin region could

Figure 4.13 An example of how unequal crossing-over could have produced the genomic regions represented in λmA14 and λmA36



have caused a duplication producing λmA14 and λmA36. The prevalence of families of repetitive elements scattered throughout the mouse genome would suggest that they might be responsible for many large block duplications. Indeed repeat sequence has been found associated with a duplication in the human genome. A short direct repeat at each end of the human ancestral foetal gene was proposed to be involved in the 5.0kb tandem duplication which occurred within the β-globin gene locus to form γ^A and γ^B (Smithies *et al.*, 1981).

Although it is easy to see the role of direct repeats in gene duplication it is curious that inverted repeats have also been found associated with duplication and amplifications in a number of cases (Fornace *et al.*, 1984; Richards *et al.*, 1983; Ford & Fried, 1986). The role of these inverted repeats in the duplication / amplification is not known but it is an intriguing alternative possibility that the arrangement of LINE members into an inverted repeat could, by a mechanism at present unknown, have led to the duplication.

An asterisk (*) indicates that a reference is missing at this position and is listed instead on page 210.

References

Allet, B., Katagiri, K.J. & Gesteland, R.F. (1973) *J. Mol. Biol.* **78**, 589 - 600.

Amor, M., Tosi, M., Duponchel, C., Steinmetz, M. & Meo, T. (1985) *Proc. Natl. Acad. Sci, U.S.A* **82**, 4453 - 4457.

Anagnou, N.P., O'Brien, S.J., Shimada, T., Nash, W.G., Chen, M. & Nienhuis, A.W. (1984) *Proc. Natl. Acad. Sci. U.S.A* **81**, 5170 - 5174.

Ananiev, E.V., Barsky, V.E., Ilyin, X., Yu, V. & Ryzic, M.V. (1984) *Chromosoma* **90**, 366 - 377.

Battey, J., Max, E., McBride, O., Swan, D. & Leder, P. (1982) *Proc. Natl. Acad. Sci. U.S.A* **79**, 5956 - 5960.

Benchimol, S., Jenkins, J.R., Crowford, L.V., Leppard, K., Lamb, P., Williamson, N.M., Pim, D.C. & Harlow, E. (1984) *Cancer Cells* **2**, 383 - 391.

Benham, F.J., Hodgkinson, S. & Davis, K.E. (1984) *EMBO J.* **3**, 2635 - 2640.

Bennett, K.L. & Hastie, N.D. (1984) *EMBO J.* **3**, 467 - 472.

Bergsma, D.J., Chang, K.S. & Schwartz, R.J. (1985) *Mol. Cell. Biol.* **5**, 1151 -1162.

Bernstein, L.B., Mount, S.M. & Weiner, A.M. (1983) *Cell* **32**, 461 - 472.

Birnboim, H.C. & Doly, J., (1979), *Nucleic Acids Res* **7**, 1513 - 1523.

Bishop, J.O., Rosbash, N. & Evans, D. (1974) *J. Mol. Biol.* **85**, 75 - 86.

Blake, C. (1983) *Nature* **306**, 535 - 537.

Blattner, F.R., Blechl, A.E., Denniston-Thomson, K., Faber, H.E., Richards, J.E., Slightom, J.L., Tucker, P.W. & Smithies, O. (1978) *Science* **202**, 1279 - 1284.

Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349 - 383.

Britten, R.J. & Kohne, D.E. (1968) *Science* **161**, 529 - 546.

Brown, S.D.M. (1983) *Gene* **23**, 95 - 97.

Brutlag, D., Appels, R., Denis, E.S. & Peacock, W.J. (1977) *J. Mol. Biol.* **112**, 31

-47.

Bucheton, A., Paro, R., Sang, H.M., Pelisson, A. & Finnegan, D.J. (1984) *Cell* **38**, 153 - 163.

Buckingham, M.E. & Minty, A.J. (1983) *In Eukaryotic Genes* (MacLean, N., Gregory, S.P. & Flavell, R.A.,eds) Chapter **21**, 365 - 397. Butterworth, London.

Buckingham, M.E. (1985) Essays in *Biochemistry* **20**, 77 - 109.

Carmon, Y., Czosnek, H., Nudel, U., Shani, M. & Yaffe, D. (1982) *Nucleic Acids Res.* **10**, 3085 - 3097.

Carroll, S.L., Bergsma, D.J. & Schwartz, R.J. (1986) *J. Biol. Chem.* **261** (19), 8965 - 8976.

Chaconas, G. & van de Sande, J.H. (1980) *Meths. Enzymol.* **65**, 75 - 85.

Challberg, M.D. & Englund, P.T. (1980) *Meths. Enzymol.* **65**, 39 - 42.

Chang, E.H., Gonda, M.A., Ellis, R.W., Scolnick, E.M. & Lowy, D.R. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4848 - 4852.

Chang, K.S., Zimmer, W.E., Jr., Bergsma, D.J., Dodgson, J.B. & Schwartz, R.J., (1984) *Mol. Cell Biol.* **4**, 2498 - 2508.

Chang, K.S., Rothblum, K.N. & Schwartz, R.J. (1985) *Nucleic Acids Res.* **13**, 1223 - 1237.

Chech, T.R. & Hearst, J.E. (1975) *Cell* **5**, 429 - 446.

Chen, M.J., Shimada, T., Moulton, A.D., Harrison, M. & Nienhuis, A.W. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 7435 - 7439.

Cheng, S.-M. & Schildkraut, C.L. (1980) *Nucleic Acids Res.* **8**, 4075 - 4090.

Church, G.M. & Gibert, W. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1991 - 1995.

Clark, A.J., Ghazal, P., Bingham, R.J. , Barrett, T. & Bishop, J.O. (1985) *EMBO J.* 4 (12), 3159 - 3165.

Clearly, M.L., Haynes, J.R., Schon, E.A. & Lingrel, J.B. (1980) *Nucleic Acid Res.* **8**, 4791 - 4802.

Clearly, M.L., Schon, E.A. & Lingrel. J.B. (1981) *Cell* **26**, 181 - 190.

Cleveland, D.W., Lopata, M.A., MacDonald, R.J., Cowan, N.J., Rutter, W.J. & Kirschner, M.W. (1980) *Cell* **20**, 95 - 105.

Cooper, E.D. & Crain, W.R., Jr. (1982) *Nucleic Acids Res.* **10**, 4081 - 4092.

Czosnek, H., Nudel, U., Shani, M., Barker, P.E., Pravtcheva, D.D., Ruddle, F.H. & Yaffe, D. (1982) *EMBO J.* **1**, 1299 - 1305.

Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M.L., Venta, P.J. & Tashian, R.E. (1982) *Nature* **298**, 297 - 300.

Czosnek, H., Nudel, U., Mayer, Y., Barker, P.E., Pravtcheva, D.D., Ruddle, F.H. & Yaffe, D. (1983) *EMBO J.* **2**, 1977 - 1979.

Dawid, I., Long, E.O., DiNocera, P.P. & Pardue, M.L. (1981) *Cell* **25**, 399 - 408.

Deininger, P.L., & Schmid, C.W. (1976) *J. Mol. Biol.* **106**, 773 - 790.

Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. & Schmid, C.W. (1981) *J. Mol. Biol.* **151**, 17 - 33.

Denison, R.A., Van Arsdell, S.W., Bernstein, L.B. & Weiner, A.W., (1981) *Proc. Natl. Acad. Sci.* **78**, 810 - 814.

Devreux, J.R., Haeberli, P. & Smithies, O. (1984) *Nucleic Acid Res.* **12**, 387 -395.

DiGiovanni, L., Haynes, S.R., Misra, R. & Jelinek, W.R. (1983) *Proc Natl. Acad. Sci. U.S.A.* **80**, 6533 - 6537.

DiNocera, P.P., Digan, M.E. & Dawid, I.B. (1983) *J. Mol. Biol.* **168**, 715 - 727.

Dowsett, A.P., & Young, M.L. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4570 -4574.

Dudov, K.P. & Perry, R.P. (1984) *Cell* **37**, 457 - 468.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Barralle, F.E., Shoulders, C.C. & Proudfoot, N.J. (1980) *Cell* **21**, 653 - 668.

Eldridge, J., Zehner, Z. & Paterson, B.M. (1985) *Gene* **36**, 55 - 63.

Engel, J.N., Gunning, P.W. & Kedes, L.H. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**,

4674 - 4678.

Erba, H:P., Gunning, P. & Kedes, L.H. (1986) *Nucleic Acids Res.* **14**, 5275 - 5294.

* Fanning, T.G. (1982) *Nucleic Acids Res.* **10**, 5003 - 5013.

Fanning, T.G. (1983) *Nucleic Acids Res.* **11**, 5073 - 5091.

Feinberg, A.P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6 - 11.

Foran, D.R., Johnston, P.J. & Moore, G.P. (1985) *J. Mol. Evol.* **22**, 108 - 116.

Ford, M. & Fried, M. (1986) *Cell* **45**, 425 - 430.

Fornace, A.J., Cummings, D.E., Comeau, C.M., Kant, J.A. & Crabtree, G.R. (1984) *Science* **224**, 161 - 164.

Fornwald, J.A., Kuncio, G., Peng, I. & Ordahl, C.P. (1982) *Nucleic Acids Res.* **10**, 3861 - 3876.

Freytag, S.O., Bock, H-G.O., Beaudet, A.L. & O'Brien, W.E. (1984) *J. Biol. Chem.* **259**, 3160 - 3166.

Fritsch, E.F., Lawn, R.N. & Maniatis, T. (1980) *Cell* **19**, 959 - 972.

Fujimoto, S., Tsuda, T., Toda, M. & Yamagishi, H. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2072 - 2076.

Fyrberg, E.A., Klindle, K.L., Davidson, N. & Sodja, A. (1980) *Cell* **19**, 365 - 378.

Fyrberg, E.A., Bond, B.J., Hershey, N.D.,Mixter, K.S. & Davidson, N. (1981) *Cell* **24**, 107 - 116.

Gabbiani, G., Schmid, E., Winter, S., Chaponnier, C., de Chastonay, C. Vandekerckhove, J., Weber, K. & Franke, W.W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 298 - 302.

Gallwitz, D. & Sures, I. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2546 - 2550.

Garner, I., Minty, A.J., Alonso, S., Barlon, P.J. & Buckingham, M.E. (1986) *EMBO J.* **5** (10), 2559 - 2567.

* Gebhard, W., Meitinger, T., Hochtl, J. & Zachau, H.G. (1982) *J. Mol. Biol.* **157**, 453 - 471.

Gebhard, W. & Zachau, H.G. (1983) *J. Mol. Biol.* **170**, 255 - 270.

Gething, M.J., Bye, J., Skelhel, J. & Waterfield, M. (1980) *Nature* **287**, 301 - 306.

Ghazal, P., Clark, A.J. & Bishop, J.O. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4182 - 4185.

Goodman, M., Koop, B.F., Czelusniak, J. , Weiss, M.J. & Slightom, J.L. (1984) *J. Mol. Biol.* **186**, 803 - 823.

Goosens, M., Dozy, A., Embury, S., Zacharides, Z., Hadjimas, M., Stamatoyannoboulos, G. & Kan, Y.W. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 518 - 521.

Green, M.R., Maniatis, T. & Melton, D.A. (1983) *Cell* **32**, 681 - 694.

Grimaldi, G., Queen, C. & Singer, M.F. (1981) *Nucleic Acids Res.* **9**, 5553 - 5568.

Grimaldi, G., Skowronski, J. & Singer, M.F. (1984) *EMBO J.* **3**, 1753 - 1759.

Gronenborn, B. & Messing, J. (1978) *Nature* **272**, 375 - 377.

Gundelfinger, E.D., Krausse, E., Melli, M. & Dobberstein, B. (1983) *Nucleic Acids Res.* **11**, 7363 - 7374.

Gundelfinger, E.D., diCarlo, M., Zoff, D. & Melli, M. (1984) *EMBO J.* **3**, 2325 - 2335.

Gunning, P., Ponte, P., Blau, H. & Kedes, L. (1983a) *Mol. Cell. Biol.* **3**, 1985 - 1985.

Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H., Kedes, L. (1983b) *Mol. Cell. Biol.* **3**, 787 - 795.

Gunning, P., Ponte, P., Kedes, L., Eddy, R. & Shows, T. (1984a) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1813 - 1817.

Gunning, P., Mohun, T., Ng, S, Y., Ponte, P. & Kedes, L. (1984b) *J. Mol. Evol.* **20** , 202 - 214.

Gwo-Shu Lee, M., Lewis, S.A., Wilde, C.D. & Cowan, N. (1983) *Cell* **33**, 477 - 478.

Halligan, B.D., Davis, J.L., Edwards, K.A. & Lin, L.F. (1982) *J. Biol. Chem.* **257**, 3995 - 4000.

Hamada, H., Petrino, M.G. & Kakunaga, T. (1982), *Proc. Natl. Acad. Sci. U.S.A.*

79, 5901 - 5905.

Hammerling, U., Ronne, H., Widmark, E., Servenius, B., Denaro, M., Rask, L. & Peterson, P.A. (1985) *EMBO J.* 4 (6). 1431 - 1434.

Hanauer, A. & Mandel, J.L. (1984) *EMBO J.* **3**, 2627 - 2633.

Hardies, S.C., Edgell, M.H. & Hutchison, C.A., III. (1984) *J. Biol. Chem* **259**, 3748 - 3756.

Hardison, R.C. (1984) *Mol. Biol. Evol.* **1**, 390 - 397.

Hardman, N. & Jack, P.L. (1977) *Eur. J. Biochem.* **74**, 275 - 283.

*
Hardman, N., Jack, P.L., Brown, A.J.P. & McLachlan, A. (1979b) *Eur. J. Biochem.* **94**, 179 - 187.

Hardman, N., Jack, P.L., Fergie, R.C. & Gerrie, L.M. (1980) *Eur. J. Biochem.* **103**, 247 - 257.

Hayashi, K., (1981) *Nucl Acids Res.* **9**, 3379 - 3388.

Haynes, S.R., Toomey, T.P., Leinwand, L. & Jelinek, W.R. (1981) *J. Mol. Cell. Biol.* **1**, 573 - 583.

Hayward, L.J. & Schwartz, R.J. (1986) *J. Cell. Biol.* **102**, 1485 - 1493.

Hattori, M., Kuhara, S., Takenaka, O. & Sakaki, Y. (1986) *Nature* **321**, 625 - 628.

Higgins, D.R., Old, J.M., Pressely, L., Clegg, J.B. & Weatherall, D. (1980) *Nature* **284**, 632- 635.

Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. & Leder, P. (1982) *Nature* **296**, 321 - 325.

Holmes, D.S. & Quigley, M. (1981) *Anal. Biochem.* **114**, 193 - 203.

*
Houck, C.M., Rinehart, F.P. & Schmid, C.W. (1979) *J. Mol. Biol.* **132**, 289 - 306.

Hu, M.C. T., Sharp, S.B. & Davidson, N. (1986) *Mol. Cell. Biol.* **6**, 15 - 25.

Hunt, J.A., Bishop, A.J.G., III. & Carson, H.L. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7146 - 7150.

Jackson, M., Heller, D. & Leinwand, L. (1985) *Nucleic Acids Res.* **13**, 3389 - 3403.

Jacq, C., Miller, J.R. & Brownlee, G.G. (1977) *Cell* **12**, 109 - 120.

Jagadeeswaran, P., Forget, B.G. & Weissman, S.M. (1981) *Cell* **26**, 141 - 142.

Jahn, C.L., Hutchison, C.A., III, Phillips, S.J., Weaver, S., Haigwood, N.L., Voliva, C.F. & Edgell, M.H. (1980) *Cell* **21**, 159 - 168.

*
Jeffreys, A J., Barrie, P.A., Harris, S., Fawsett, D.M. , Nugent, Z.J. & Boyd, A.C. (1982) *J. Mol. Biol.* **156**, 487 - 503.

*
Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, G.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. & Schmid, C.W. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1398 - 1402.

Jelinek, W.R. & Schmid, C.W. (1982) *Annu. Rev. Biochem.* **51**, 813 - 844.

*
Jones, R.S. & Potter, S.S. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1989 - 1993.

Junakovic, N., Caneva, R. & Ballano, P. (1984) *Chromosoma* **90**, 378 - 382.

Kalb, V.F., Glasser, S., King, D. & Lingrel, J.B. (1983) *Nucleic Acids Res.* **11**, 2177 - 2184.

Karin, M. & Richards, R.I. (1982) *Nature* **299**, 797 - 802.

Karn, J., Brenner, S., Barnett, L. & Cesareni, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77** (9), 5172 - 5176.

Katzir, N., Rechain, G., Cohen, J.B., Unger, T., Simoni, F., Segal, S., Cohen, D. & Givol, D. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1054 - 1058.

Kingsbury, D.T. (1969) *J. Bacteriol.* **98**, 1400 - 1410.

Klein, A. & Meynhas, O. (1984) *Nucleic Acids Res.* **12**, 3763 - 3776.

Kole, L.B., Haynes, S.R. & Jelinek, W.R. (1983) *J. Mol. Biol.* **165**, 257 - 286.

Kost, T.A., Theodorakis, N. & Hughes, S.H. (1983) *Nucleic Acids Res.* **11**, 8287 - 8301.

Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. & Georgiev, G.P. (1980) *Nucleic Acids Res.* **6**, 1201 - 1215.

Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P., Scryabin, K.G., Bayer, A.A & Georgiev, G.P. (1982) *Nucleic Acids Res.* **10**, 7461 - 7475.

Lacy, E & Maniatis, T. (1980) *Cell* **21**, 545 - 553.

Laird, C.D. (1971) *Chromosoma* **32**, 378 - 406.

Langer, P.R., Waldrop, A.A. & Ward, D.C. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 6633 - 6637.

Lauer, J., Shen, C.K.J. & Maniatis, T. (1980) *Cell* **20**, 119 - 130.

Leader, D.P., Gall, I. & Lehrach, H. (1985) *Gene* **36**, 369 - 374.

Leader, D.P., Gall, I., Campbell, P. & Frischauf, A.M. (1986) *DNA* **5** (3), 235 - 238.

Lee, M.G., Lewis, S.A., Wilde, C.D. & Cowan, N.J. (1983) *Cell* **33**, 477 - 487.

Leder, A., Swan, D., Ruddle, F., D'Eustachio, P. & Leder, P. (1981) *Nature* **293**, 196 - 200.

Leibhaber, S.A., Gossens, M. & Kan, Y.W. (1981) *Nature* **290**, 26 - 29.

Lemischka, I. & Sharp, P.A. (1982) *Nature* **300**, 330 - 335.

Lerman, M.I., Thayer, R.E. & Singer, M.F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80** 3966- 3970.

Leube, R. & Gallwitz, D. (1986) *Nucleic Acids Res.* **14** (15), 6339 - 6349.

Li, W.H., Gojobori, T. & Nei, M. (1981) *Nature* **292**, 237 - 239.

Liebermann, D., Hoffman-Lieberman, B., Weinthal, J., Childs, G., Maxson, R., Mauron, A., Cohen, S.N. & Kedes, L. (1983) *Nature* **306**, 342 - 347.

Limbach, K.J. & Wu, R. (1985) *Nucleic Acids Res.* **13**, 617 - 630.

Little, P.F.R. (1982) *Cell* **28**, 683 - 684.

Loeb, D.D., Padgett, R.W., Hardies, S.C., Sheshe, W.R., Comer, M.B., Edgell, M.H. & Hutchison III, C.A. (1986) *Mol. Cell. Biol.* **6**, 168 - 182.

Loening, U.E. (1967) *Biochem J,* **102**, 251 - 257.

Lund, E. & Dahlberg, J.E. (1984) *J. Biol Chem.* **259**, 2013 - 2021.

McGinnis, W., Shemoen, A.W. & Beckendorf, S.K. (1983) *Cell* **34**, 75 - 84.

McGrath, J.P., Capon, D.J., Smith, D.H., Chen, E.Y., Seeburg, P.H., Goeddel, D.V. & Levinson, A.D. (1983) *Nature* **304**, 501 - 506.

McKeown, M. & Firtel, R.A. (1981) *Cell* **24**, 799 - 807.

*
Manser, T. & Gesteland, R.F., (1981) *J. Mol. Genet.* **1**, 117 - 120.

*
Manuelidis, L. & Biro, P.A. (1982) *Nucleic Acids Res.* **10**, 3221 - 3239.

Manuelidis, L. (1982) *Nucleic Acids Res.* **10**, 3211 - 3219.

Martin, S.L., Vincent, K.A. & Wilson, A.L. (1983) *J. Mol. Biol.* **164**, 513 - 528.

Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H. & Hutchinson C.A., III, (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2308 - 2312.

Masters, J.N., Yang, J.K., Cellini, A. & Attardi, G. (1983) *J. Mol. Biol.* **167**, 23 - 36.

Maxam, A.M. & Gilbert, W. (1980) *Meth. Enzymol.* **65**, 499 - 560.

Mayer, Y., Czosnek, H., Zeelon, P.E., Yaffe, D. & Nudel, U. (1984) *Nucleic Acids Res.* **12**, 1087 - 1100.

Messing, J. (1981) *Proceedings of the Third Cleveland Symposium on Macromolecules*, Elsevier, Amsterdam, 143 - 153.

*
Miller, J.R., Cartwright, E.M., Brownlee, G.G., Fedoroff, N.V. & Brown, D. (1978) *Cell* **13**, 717 - 725.

Miller, J.R. & Melton, D.A. (1981) *Cell* **24**, 829 - 835.

Minty, A.J., Caravatti, M., Robert, B., Cohen, A., Daubas, P., Weydert, A., Gros, F. & Buckingham, M.E. (1981) *J. Biol. Chem.* **256**, 1008 - 1014.

Minty, A.J., Alonso, S., Guenet, J.L. & Buckingham, M.E. (1983) *J. Mol. Biol.* **167**, 77 - 101.

Miyata, T. & Hayashida, H. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5739 - 5743.

Miyata, T. & Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 450 - 453.

Miyoshi, J., Kagimoto, M., Soeda, E. & Sakaki, Y. (1984) *Nucleic Acids Res.* **12**, 1821 - 1828.

Modelell, J. , Bender, W. & Meselson, M. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1678 - 1672.

Monstein, H-J., Hammerstrom, K., Westin, G., Zabielski, J., Philipson, L. & Pettersson, U. (1983) *J. Mol. Biol.* **167**, 245 - 252.

Moos, M. & Gallwitz, D. (1982) *Nucleic Acids Res.* **10**, 7843 - 7849.

Moos, M. & Gallwitz, D. (1983) *EMBO J.* **2**, 757 - 761.

Mottez, E., Rogan, P.K. & Manuelidis, L. (1986) *Nucleic Acids. Res.* **14** (7), 3119 - 3136.

Nakajima-Iijima, S., Hamada, H., Reddy, P. & Kakunaga, T. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6133 - 6137.

Ng, R. & Abelson, J. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3912 - 3916.

Ng, S.Y., Gunning, P., Eddy, R., Ponte, P., Leavitt, J., Shows, T. & Kedes, L. (1985) *Mol. Cell. Biol.* **5**, 2720 - 2732.

Nishioka, Y., Leder,A. & Leder, P. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2806 - 2809.

Notake, M., Tobimastsu, T., Watanabe, Y., Takahashi, H., Mishina, M. & Numa, S. (1983) *FEBS Lett.* **156**, 67 - 71.

Nudel, U., Zakut, R., Katcoff, D. Carmon, Y., Czosnek, H., Shani, M. & Yaffe, D. (1982a) *In Muscle Development,* Pearson, M.L. & Epstein, E.F., Eds., 177 - 188, Cold Spring Harbor Laboratory, N.Y.

Nudel, U., Katcoff, D., Zakut, R., Shani, M., Carmon, Y., Finer, M., Czosnek, H., Ginsburg, I. & Yaffe, D. (1982b) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2763 - 2767.

Nudel, U., Zakut, R., Shani, M., Neuman, S., Levy, Z. & Yaffe, D. (1983) *Nucleic Acids Res.* **11**, 1759 - 1771.

O'Farrell, P. (1981) *Focus* **3**, 1 - 8.

* Ordahl, C.P. & Cooper, T.A. (1983) *Nature* **303**, 348 - 349.

Orkin, S.H., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D.J. & Nathans, D.G. (1979) *Cell* **17**, 33 - 42.

* Perler, F., Efstrtiadis, A., Lomedico, P., Gilbert, W., Kolodnev, R. & Dodgson, J. (1980) *Cell* **20**, 555 - 566.

Perlman, S., Phillips, C.A. & Bishop, J.D. (1976) *Cell* **8**, 33 - 42.

Perled-Yalif, E., Cohen-Binder, I. & Meynhas, O. (1984) *Gene* **29**, 157 - 166.

Pichauntes, S., Medina, A., Bell, G., Gomez, I., Valenzuela, P., Bull, P .& Venegas, A. (1982) *Arch. Biol. Med. Exp.* **15**, 381 - 394.

Piechaczyk, M., Leyal-Taha, M.N., Sri-Widada, J., Brunel, C., Phiautard, J. & Jeanteur, P., (1982) *Nucleic Acids Res.* **10**, 4627 - 4640.

Ponte, P., Gunning, P., Blau, H. & Kedes, L. (1983) *Mol. Cell. Biol.* **3**, 1783 - 1791.

Ponte, P., Ng, S.-Y., Engel, J., Gunning, P. & Kedes, L. (1984) *Nucleic Acids Res.* **12** (3), 1687 - 1696.

Popp, R.A., Lalley, P.A., Whitney, J.B. & Anderson, W.F. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 6362 - 6366.

Potter, S.S., Truett, M., Phillips, M. & Maher, A. (1980) *Cell* **20**, 639 - 647.

Potter, S.S. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1012 - 1016.

Proudfoot, N.J. (1980) *Nature* **286**, 840 - 841.

Proudfoot, N.J. & Maniatis, T. (1980) *Cell* **21**, 537 - 544.

Proudfoot, N.J., Gill, A. & Maniatis, T. (1982) *Cell* **31**, 553 - 563.

Rackwitz, H-R., Zehnter, G., Frischauf, A. & Lehrach, H. (1984) *Gene* **30**, 195 - 200.

Ricca, G.A., Taylor, J.M. & Kalinyak, J.E. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 724 - 728.

Richards, J.E., Gilliam, A.C., Shen, A., Tucker, P.W. & Blattner, F.R. (1983) *Nature* **306**, 483 - 487.

Rigby, P.W.J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237 - 251.

Robert, B., Daubas, P., Akimenko, M. -A., Cohen, A., Garner, I., Guenet, J.-L. & Buckingham, M.E. (1984) *Cell* **39**, 129 - 140.

Robert, B., Barton, P., Minty, A.J., Daubas, P., Weydert, A., Bonhomme, F., Catalan, J., Chazottes, D., Guenet, J.-L. & Buckingham, M.E. (1985) *Nature* **314**, 181 - 183.

Rogers, J. (1983) *Nature* **306**, 113 - 114.

Rogers, J.H. (1984) *Int. Rev. Cytol.* **93**, 187 - 279.

Rogers, J.H. (1985) *Int. Rev. Cytol.* **93**, 187 - 279.

Roychoudhury, R. & Wu, R. (1980) *Meths. Enzymol.* **65**, 43 - 62.

Rubin, G.M., Brorein, W.J., Jr., Dunsmuir, P., Flavell, A.J., Levis, K., Strobel, E., Toole, J.J. & Young, E. (1981) *Cold Spring Harbor Symp. Quant. Biol.* **45**, 619 - 628.

Rubin, G.M., Kidwell, M.G. & Bingham, P.M. (1982) *Cell* **30**, 987 - 994.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 - 5467.

Sanger, F. & Coulson, A.R. (1978) *FEBS Lett,*. **87**, 107 - 110.

Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. & Roe, B. (1980) *J. Mol. Biol.* **143**, 161 - 178.

Scarpulla, R.C. & Wu, R. (1983) *Cell* **32**, 473 - 482.

Scarpulla, R.C. (1984) *Mol. Cell. Biol.* **4**, 2279 - 2288.

Scheller, R.H., McAllister, L.B., Crain,W.R., Jr., Durica, D.S., Posakony, J.W., Thomas, T.L., Britten, R.J. & Davidson, E.H. (1981) *Mol. Cell. Biol.* **1**, 609 - 628.

Scherer, G., Tshudi, C., Perera, J., Deluis, H. & Pirrotta, V. (1982) *J. Mol. Biol.* **157**, 435 - 451.

Schindler, C.W. & Rush, M.G. (1985) *J. Mol. Biol.* **181**, 161 - 173.

Schmid, C.W., Manning, J.E. & Davidson, N. (1975) *Cell* **5**, 159 - 172.

Schmid, C.W. & Deininger, P.L. (1975) *Cell* **6**, 345 - 358.

Schmid, C.W. & Jelinek, W.R. (1982) *Science* **216**, 1065 - 1070.

Schon, E.A., Wernke, S.M. & Lingrel, J.B. (1982) *J. Biol. Chem.* **257**, 6825 - 6835.

Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. & Maio, J.J. (1983) *Nature* **304**, 277 - 280.

Sharp, P.A. (1983) *Nature* **301**, 471 - 472.

Shen, S., Slightom, J.L. & Smithies, O. (1981) *Cell* **26**, 191 - 203.

Shimada, T., Chen, M.J. & Nienhuis, A.W. (1984) *Gene* **31**, 1 - 8

Shyman, S. & Weaver, S. (1985) *Nucleic Acids Res.* **13** (14), 5085 - 5093.

Singer, M.F. (1982) *Cell* **28**, 433 - 434.

Singer, M.F., Thayer, K.E., Grimaldi, G., Lerman, M.I. & Fanning, T.G. (1983) *Nucleic Acids Res.* **11**, 5739 - 5745.

Singer, M.F. & Skowronski, J. (1985) *Trends. Biochem. Sci.* **10**, 119 - 122.

Skowronski, J. & Singer, M.F. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6050 - 6054.

Slightom, J.L., Blechl, A.E. & Smithies, O. (1980) *Cell* **21**, 627 - 638.

Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L. & Snen, S. (1981) *Cell* **26**, 345 - 353.

Soares, M.B., Schon, E. & Efstratiadis, A. (1985a) *J. Mol. Evol.* **22**, 117 - 133.

Soares, M.B., Schon, E., Henderson, A., Cate, K.R., Zeitlin, S., Chirgwin, J. & Efstratiadis, A. (1985b) *Mol. Cell. Biol.* **5**, 2090 - 2103.

Soriano, P., Szabo, P. & Bernardi, G. (1982) *EMBO J.* **1** (5), 579 - 583.

Soriano, P., Meunier-Rotival, M. & Bernardi, G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1816 - 1820.

Southern, E.M. (1975) *J. Mol. Biol.* **94**, 51 - 69.

Southern, E.M. (1975) *J. Mol. Biol.* **98**, 503 - 517.

Spradling, A.C. & Rubin, G.M. (1981) *Annu. Rev. Genetics* **15**, 219 - 264.

Spritz, R.A., DeRiel, J.K., Forget, B.G. & Weissman, S.M. (1980) *Cell* **21**, 639 - 646.

Staden, R. (1977) *Nucleic Acids Res.* **4**, 4037 - 4051.

Staden, R. (1979) *Nucleic Acids Res.* **6**, 2601 - 2610.

Stein, J. P., Munjaal, R.P., Lagace, L., Chai, E., O'Malley, B.W. & Means, A.R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6485 - 6489.

Sutcliffe, J.G. (1978) *Nucleic Acids Res.* **5**, 2721 - 2728.

Sutcliffe, J.G., Milner, R.J., Bloom, F.E. & Lerner, R.A. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4942 - 4946.

Ueda, S., Nakai, S., Nishida, Y., Hisajima, H. & Honjo, T. (1982) *EMBO J.* **1**, 1539 - 1544.

Ueyama, H., Hamada, H., Battula, N. & Kakunaga, T. (1984) *Mol. Cell. Biol.* **4**, 1073 - 1078.

Uhler, M., Herbert, E., D'Eustachio, P. & Ruddle, F.D. (1983) *J. Biol. Chem* **258**, 9444 - 9453.

Ullu, E. & Tschudi, C. (1984) *Nature* **312**, 171 - 172.

Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. & Gesteland, R.F. (1981) *Cell* **26**, 11 - 17.

Van Arsdell, S.W. & Weiner, A.M. (1984) *Nucleic Acids Res.* **12**, 1463 - 1471.

Vandekerckhove, J. & Weber, K. (1978a) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1106 - 1110.

Vandekerckhove, J. & Weber, K. (1978b) *J. Mol. Biol.* **126**, 783 - 802.

Vandekerckhove, J. & Weber, K. (1979a) *Differentiation* **14**, 123 - 133.

Vandekerckhove, J. & Weber, K. (1979b) *FEBS Lett.* **102**, 219 - 222

Vandekerckhove, J., Franke, W.W. & Weber, K. (1981) *J. Mol. Biol.* **152**, 413 - 426.

Vandekerckhove, J., de Couet, H.-G. & Weber, K. (1983) *Academic Press* Sydney, Australia, 241 - 248.

Vandekerckhove, J. & Weber, K. (1984) *J. Mol. Biol.* **179**, 391 - 413.

Vanin, E.F., Goldberg, G.I., Tucker, P.W. & Smithies, O. (1980) *Nature* **286**, 222 - 226.

Vanin,E.F. (1983) *Regulation of Hemoglobin Biosynthesis* ed. Goldwasser, E. **21**, 69 - 88.

Vanin, E.F. (1984) *Biochem. Biophys. Acta.* **782**, 231 - 241.

Vanin, E.F. (1985) *Ann. Rev. Genet.* **19**, 253 - 272.

Varshney, U. & Gedamu, L. (1984) *Gene* **31**, 135 - 145.

Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A.,III. & Edgell, M.H. (1983)

Nucleic Acids Res. **11**, 8847 - 8859.

Voliva, C.F., Martin, S.L., Hutchison, C.A., III. & Edgell, M.H. (1984) *J. Mol. Biol.* **178**, 795 - 813.

Walter, P. & Blobel, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7112 - 7116.

Weaver, S., Comer, M.B., Jahn, C.L., Hutchison, C.A., III & Edgell, M.H. (1981) *Cell* **24**, 403 - 411.

Weissbach, A. (1977) *Annu. Rev. Biochem.* **46**, 25 - 47.

Westin, G., Monstein, H-J., Zabielski, J., Phiipson, L. & Pettersson, U., (1981) *Nucleic Acids Res.* **9**, 6323 - 6338.

Weydert, A., Daubas, P., Caravatti, M., Minty, A., Bugaisky, G., Cohen, A., Robert, B. & Buckingham, M. (1983) *J. Biol. Chem* **258**, 13867 - 13881.

Wiedemann, L.M. & Perry, R.P. (1984) *Mol. Cell. Biol.* **4**, 2518 - 2528.

Wilde, C.D., Crowther, C.E., Cripe, T.P., Lee, M. G-S. & Cowan, N.J. (1982a) *Nature* **297**, 83 - 84.

Wilde, C.D., Crowther, C.E. & Cowan, N.J. (1982b) *Science* **217**, 549 - 550.

Wilson, D.A. & Thomas, C.A., Jr. (1974) *J. Mol. Biol.* **84**, 115 - 144.

Wilson, R. & Storb, U. (1983) *Nucleic Acids Res.* **11**, 1803 - 1816.

Witney, F.R. & Furano, A.V. (1984) *J. Biol Chem.* **259**, 10481 - 10492.

Yaffe, D., Nudel, U., Mayer, Y. & Neuman, S. (1985) *Nucleic Acids Res.* **13**, 3732 - 3737.

Yanisch-Perron, C., Vieira, J. & Messing, J. (1985) *Gene* **33**, 103 - 119.

Young, M.W. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 6274 - 6278.

Zabarovsky, E.R., Chumakov, I.M., Prassolov, V.S. & Kisselev, L.I. (1984) *Gene* **30**, 107 - 111.

Zakut, R., Shani, M., Givol, D., Neuman, S., Yaffe, D. & Nudel, U. (1982) *Nature* **298**, 857 - 859.

Zakut-Houri, R., Oren, M., Bienz, B., Lavie, V., Hazum, S. & Givol, D. (1983) *Nature* **306**, 594 - 597.

Adams, R. L. P., Burdon, R. H. & Fulton, J. (1983) *Biochem. Biophys. Res. Commun.* **113**, 695 - 702.

Antoine, M & Niessing, J. (1984) *Nature* **310**, 795 - 798.

Bostock, C. (1980) *Trends Biochem. Sci.* **5**, 117 - 119.

Collins, J. & Elzinga, M. (1975) *J. Biol. Chem.* **250**, 5915 - 5920.

Evans, H. J., Gosden, J. R., Mitchell, A. R. & Buckland, R. A. (1974) *Nature* **251**, 346 - 347.

Gautier, F., Bunemann, H. & Grotjahn, L. (1977) *Eur. J. Biochem.* **80**, 175 - 183.

Hardman, N., Bell, A.J. & McLachlan, A. (1979a) *Biochim. Biophys. Acta* **564**, 372 - 389.

Horz, W. & Altenburger, W. (1981) *Nucl. Acids Res.* **9**, 683 - 696.

Jamrich, M., Warrior, R., Steele, R. & Gall, J. G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3364 - 3367.

Jelinek, W. R. (1978) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2679 - 2683.

Jones, K. W. & Corneo, G. (1971) *Nature (London) New Biol.* **233**, 268 - 271.

Lu, R. & Elzinga, M. (1977) *Biochemistry* **16**, 5801 - 5806.

MacLeod, A.R. & Talbot, K. (1983) *J. Mol. Biol* **167**, 523 - 537.

Maio, J. J., Brown, F. L. & Musich, P. R. (1977) *J. Mol. Biol.* **117**, 637 - 655.

Manuelidis, L. (1978) *Chromosoma* **66**, 1 - 21.

Miklos, G. L. G. & John, B. (1979), *Amer. J. Human Genet.*, **31**, 264 - 280.

Ohno. S. (1971) *Nature*, **234**, 134 - 137.

Pech, M., Streeck, R. E. & Zachau, H. G. (1979) *Cell* **18**, 883 - 893.

Taparowsky, E. J. & Gerbi, S. A. (1982) *Nucl. Acids Res.* **10**, 5503 - 5515.

Vandekerckhove, J. & Weber, K. (1978c) *Eur. J. Biochem.* **90**, 451 - 462.

Varley, J. M., MacGregor, H. C. & Erba, H. P. (1980) *Nature* **283**, 686 - 688.